THE UNIVERSITY *of* EDINBURGH

Modelling subphonemic information flow: An investigation and extension of Dell's (1986) model of word production

Helen Susannah Moat, MSci., M.Sc.

A thesis submitted in fulfilment of requirements for the degree of Doctor of Philosophy

to

School of Philosophy, Psychology and Language Sciences
University of Edinburgh

September 2010

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Helen Susannah Moat

Abstract

Dell (1986) presented a spreading activation model which accounted for a number of early speech error results, including the relative proportions of anticipations, perseverations and exchanges found in speech error corpora, the lexical bias effect, the phonological similarity effect, and the effect of speech rate on error rate. This model has had an immense influence on the past 20 years of research into word production, with the original paper being cited over 1,000 times.

Many studies have questioned how activation should flow between words and phonemes in this model. This thesis aimed to clarify what current speech error evidence tells us about how activation flows between phonemes and subphonemic representations, like features. Does activation cascade from phonemes to features, and does it feed back? The work presented here extends previous modelling investigations in two ways. Firstly, whereas previous modelling research has tended to evaluate model behaviour using arbitrarily chosen parameter settings, we illuminate the influence of the parameters on model behaviour and propose methods to draw general conclusions about model behaviour from large numbers of simulations at orthogonally varied parameter settings. Secondly, we extend the scope of the simulations to consider output at a subphonemic level, modelling recent data acquired via acoustic and articulatory measurements, such as voicing onset time (VOT), electropalatography (EPG) and ultrasound, alongside older transcribed speech error data. Throughout the thesis, we consider whether parameter settings which lead the model to capture individual results also permit other results to be accounted for and do not cause otherwise implausible behaviour.

Through manipulating parameter settings in Dell's (1986) original model, we find that increasing the number of steps before selection generally does not decrease the error rate, but rather increases it, contrary to results reported by Dell (1986). This calls into question the claim that an increase in steps before selection provides a good model of a slower speech rate. We also demonstrate that the model captures the negative correlation reported by Dell, Burger, and Svec (1997) between error

rate and the ratio of anticipations to perseverations, and further predicts that there should be a negative correlation between this ratio and the proportion of errors which are non-contextual. However, our results show that no parameter setting allows the model to generate enough exchanges to match even minimum estimates from a reanalysis of multiple speech error corpus reports, without falling foul of other constraints; in particular, limits on the overall number of errors generated. We suggest that the exchange completion triggering mechanism proposed by Dell (1986) is not strong enough, and that current corpus evidence provides little support for his account of word sequencing.

Focusing on single word production therefore, the second part of the thesis investigates behaviour of models with output at a subphonemic level. We find that, provided sufficient contextual errors occur at the featural level, a model in which only the identity of the selected phoneme is conveyed to the featural level can account for: (i) the phonological similarity effect found in transcribed records of speech errors (whereas in models with output at the phoneme level, feedback from features to phonemes is required); (ii) detectable influences of intended phonemes in VOT measurements of unintended phonemes, as well as the effect of error outcome lexicality on these results (findings presented in support of cascading from phonemes by Goldrick & Blumstein, 2006); and (iii) increased similarity of EPG measurements of articulations to reference measurements of competing articulations when production of the competing onset would result in a word (McMillan, Corley, & Lickley, 2009). Initial results appear to confirm however that, in contrast, phonological similarity effects on the relationship of articulatory and acoustic measurements of productions to reference measurements (McMillan, 2008) can only be accounted for in an architecture with feedback from features to phonemes. To strengthen conclusions about articulatory evidence of lexical bias and phonological similarity effects, future work needs to consider the extremely strong effects of frequency observed in these simulations.

The results presented in this thesis contribute to a greater comprehension of the behaviour of Dell's (1986) influential model, and further demonstrate that the model can be extended to account for new instrumental evidence, whilst clarifying the constraints on activation flow between phonemes and features which this new evidence imposes.

Acknowledgements

I owe thanks to a number of organisations and individuals for their support while I have been working on this thesis.

Having sparked my initial interest in word production, Martin Corley has continued to provide advice and practical support throughout my PhD. I am particularly grateful to him for his wise counsel on the subject of academic writing, and his endless encouragement and belief in this project.

I am also indebted to Rob Hartsuiker, especially for allowing me to visit his lab in Ghent. His suggestions had an important influence on the work reported here.

The members of the Edinburgh Disfluency Group have been a source of both academic discussions and deeply appreciated office banter. I owe particular thanks to Corey McMillan, whose enthusiasm and research were a notable source of inspiration in my early PhD years. I am also grateful to Paul Brocklehurst, Phil Collard, Lucy MacGregor, Michael Schnadt, and Ollie Stewart, for much support and laughter. Paul and Ollie helped out with proof reading, and Ollie deserves particular mention for the combination of practical assistance and humour he has provided over the last few months. Thanks also to Dina van der Hulst for conversations about cricket.

This project would have been much more difficult without the support provided by Katie Keltie, Toni Noble, Kirsty Woomble, Steven McGauley, and Davy Wilkinson. I owe further thanks to John Blair-Fish and Mike Baker for their willingness to advise and assist with use of the Edinburgh Compute and Data Facility (ECDF), which made this project possible. I am forever indebted to my one and only participant, Eddie, for his inexhaustible enthusiasm for taking part in multiple simultaneous experiments deep into the night.

Funding for this work was provided by the Economic and Social Research Council (grant PTA-031-2004-00279), and travel funding in particular was supplemented by The School of Philosophy, Psychology and Language Sciences at the University of

Edinburgh and a Grindley grant. I am grateful for the comments on this research which were offered at a number of conferences, but the feedback I received from reviewers at CogSci 2008 was of particular use in developing my ideas further.

Extremely heartfelt thanks go to Zeynep Ilkin, who has been a constant source of friendship, practical support and psycholinguistic inspiration throughout my PhD. Immense gratitude also goes to all the drummers and dancers of The Edinburgh Samba School, who have served as chief guardians of my sanity for the past few years. While many of them deserve to be named here, I am particularly grateful to my *velho amigo*, Mestre Allysson Velez, for the inspiration he has provided from near and afar, and to my partner in hatted crime, Lindsay Hunter, for her greatly valued friendship. Bruna Werneck has provided huge amounts of support with an admirable disregard for geography, and Andy Chung still can't spell dusty. Dave Murray-Rust is owed whisky to thank him for food, interesting questions and much more. Thanks to Chris Elliott for wonderful Italian intermissions, and sincere apologies to Paul Harrison for shattering his teenage hopes that I would never become boring enough to write a PhD.

Finally, I owe huge thanks to my brilliant family, the Moats, the Whittinghams, the Woogaras and the Brightmans, for the endless supply of phone calls, encouragement, love and food, and a much appreciated proof reading marathon. My sister Rachael has been on hand throughout with big sisterly advice and irrepressible giggles. My deepest thanks go to my Mum and Dad, Anthea and David, proper acknowledgement of whose help would require another thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> The Leith police dismisseth us
> They thought we sought to stay;
> The Leith police dismisseth us
> They thought we'd stay all day.
> The Leith police dismisseth us,
> We both sighed sighs apiece;
> And the sighs that we sighed as we said goodbye
> Were the size of the Leith police.

Anyone who has tried to recite a tongue twister such as the one above will be aware that human speech is subject to errors, and that those errors tend to be messy; for example, it is quite usual to pronounce the 'ss' in *dismisseth* as something part-way between an 's' and a 'th'. It may seem somewhat surprising, then, that the dominant implemented model of errors in speech production (Dell, 1986) only produces mistakes which are *well-formed*; that is, it allows for the fact that a 'th' may unintentionally replace an 's', but not for the fact that the result may be a blend of the two.

This assumption in this very influential model stems from a reliance on results acquired through the transcription of speech errors, which are unavoidably influenced by a human predisposition to hear whole sounds. The present thesis is written in the context of increasing evidence from new experiments involving acoustic analysis of speech sounds and recordings of articulatory movements, which show that 'blended' speech errors are common. The realisation that errors are not all speech sound sized raises new questions, which this new high resolution speech production data allows us to address. In the present thesis, we simulate acoustic and articulatory results to investigate how information in the word production system flows

between representations of speech sounds and abstract articulatory representations, which specify, for example, which part of the tongue is raised during production of a sound, or timings between a tongue movement and vibration of the vocal cords.

Investigating different models of information flow between speech sounds and articulatory characteristics is complicated, however, by the existence of eight essentially free parameters in Dell's (1986) model. Different models of information flow may well rely on different specifics of the implementation to account for empirical evidence, and these different elements may require different parameter settings in order to operate effectively. To follow the common approach of testing model behaviour at one arbitrarily chosen set of parameter settings would therefore not seem appropriate. Furthermore, there is a need for a better understanding of the influence that these parameters have on Dell's (1986) model's behaviour, to illuminate the role that parameter settings play in the ability of the model to account for different results. A better understanding of the effects of these parameters will also allow us to better understand the general properties of the model itself.

The present thesis therefore has two main aims. Firstly, we aim to clarify the influences of the free parameters in Dell's (1986) model on basic and more complex model behaviour, and to propose methods to draw general conclusions about model behaviour from large numbers of simulations at orthogonally varied parameter settings. Secondly, we aim to determine the constraints imposed by new evidence on models of information flow between speech sounds (phonemes) and articulatory characteristics (features), by extending the model to consider output at a featural level and by simulating acoustic and articulatory measures within the framework of Dell's (1986) model.

1.1 Thesis overview

The thesis is organised as follows.

Chapter 2 surveys the existing literature, summarising accounts of basic types of errors, and considering what basic errors have told us about how information flows between representations for words and phonemes. We highlight problems with old evidence relying on transcription which has been used to argue that errors are well-formed, and review claims that empirical investigators have made about information flow between phonemes and features on the basis of newer acoustic analyses of sounds and recordings of articulatory movements. Finally, chapter 2

draws attention to problems in investigating different models of information flow within Dell's (1986) architectures given the number of free parameters in the model, and argues that there is a need to clarify the influence of these parameter settings on model behaviour and to determine the role of these settings in the model's ability to account for empirical evidence.

Chapter 3 specifies the structure of the model used throughout the thesis. In this structure, a layer of words is connected to a layer of phonemes, which in turn is connected to a layer of features. We outline how processing in this structure would operate for a model which generated well-formed errors only by producing output at the phoneme level, and a new model with output at the featural level. We explain the interpretation of the output of the model, including how we simulate new acoustic and articulatory evidence. We describe the different models of information flow between words, phonemes and features which we consider in this thesis. A large scale modelling methodology to address the problem of the many free parameters in the model is then introduced.

The first simulations reported in this thesis, in chapters 4 to 6, investigate the behaviour of Dell's (1986) original model. Results are considered across a wide range of parameter settings, to clarify the influence of the parameter settings on the behaviour of the model, and to illuminate the parameter independent properties of the underlying architecture.

Chapter 4 describes the effects of manipulating the free parameters on the basic behaviour of the original model. A methodology to facilitate analysis of parameter effects is introduced. Using this methodology, we show that a parameter manipulation previously thought to correspond to how long the speaker has to prepare their production may be better conceptualised as how long the speaker has to remember what the intended message of the utterance is. Referring to previous empirical evidence, we establish upper limits on error rate and the number of errors originating outside the current phrase for a model to reasonably reflect basic human speech behaviour, and report which parameter settings allow the model to meet these constraints.

Chapter 5 considers basic errors in which phonemes move between words. We re-evaluate evidence of the relative frequency of errors in which phonemes are copied to earlier words (anticipations), phonemes are copied to later words (perseverations), and phonemes swap between words (exchanges). We demonstrate that manipulations of parameter settings in the original model allow it to account for the finding

that speakers who make more errors also make a higher proportion of perseverations (Dell, Burger, & Svec, 1997). Furthermore, we show that the model predicts that speakers who make a higher proportion of perseverations should also make more errors in which the source of the error is outside the current utterance. However, we find that parameter settings which do not cause the model to generate too many errors lead the model to generate a much lower proportion of exchanges than the proportion witnessed in humans. We conclude that the mechanism proposed by Dell (1986) to account for exchanges is deficient, and consider single word productions only for the rest of the thesis.

Chapter 6 presents a methodology for determining whether a model can account for a given statistical difference demonstrated in human behaviour, when statistical tests of the model are carried out at many different parameter settings, such that there is a very high chance of some false positive results occurring. As test cases, we use the classic results that errors which result in words are more likely to occur than errors which do not result in words (the lexical bias effect), and that phonemes which are similar are more likely to swap than phonemes which are not similar (the phonological similarity effect). Using the methodology developed in this chapter, we present statistical evidence that in a model which outputs phoneme sized units, information must flow backwards from phonemes to words in order to account for the lexical bias effect, and similarly, information must flow backwards from features to phonemes to account for the phonological similarity effect. We confirm that a model with both types of backwards information flow can account for both results, without needing to use different parameter settings for the different effects. Finally, we continue to use the methodology developed in chapter 4 to clarify which parameter settings are required for the two effects to be observed.

Chapters 7 and 8 investigate the behaviour of a model with output at the featural level, presenting the first simulations of acoustic and articulatory evidence within the framework of Dell's (1986) model, and examining the constraints that old transcribed speech error evidence and new acoustic and articulatory evidence place on models of information flow between phonemes and features.

In chapter 7, we use the methodology developed in chapter 6 to show that while it is still required that information flows backwards from phonemes to words for a model with output at the featural level to account for the classic lexical bias effect, only the identity of the phoneme chosen for production needs to be conveyed to the featural level to account for the phonological similarity effect. No information from phonemes which were not selected needs to pass to the featural level, and there

is no requirement for information to flow back from features to phonemes. This very simple model is also shown to account for instrumentally measured acoustic evidence demonstrating that phonemes produced by mistake bear characteristics of the intended phoneme (Goldrick & Blumstein, 2006). Finally, we show that this model can also explain findings that traces of an intended phoneme on an error are weaker when the error results in a word (Goldrick & Blumstein, 2006). Throughout this chapter, we note which parameter settings are required to account for different effects. We find that different models of information flow have different parameter setting requirements in order to account for Goldrick and Blumstein's (2006) evidence, validating our original concerns about testing different models of information flow at one arbitrarily chosen set of parameter settings.

Chapter 8 extends the work in the previous chapter by simulating articulatory evidence of tongue movements. We consider evidence which does not rely on categorisation of productions as erroneous or correct. Instead, all acoustic and articulatory recordings are compared to an ideal recording using the *delta method* (McMillan et al., 2009). Simulations show that the simple model in which only the identity of the phoneme chosen for production is conveyed to the featural level can account for evidence that articulations of a phoneme are more like ideal articulations of a phoneme from a nearby word if accidental production of the other phoneme would result in a word (McMillan et al., 2009). However, McMillan (2008) reports that articulations of a phoneme are less like ideal articulations of that phoneme when there is a phoneme in a nearby word which shares a high number of features with the intended phoneme, and we show that an account of this finding requires that information must flow back from features to phonemes. Conclusions in this chapter are limited by the finding that the frequency of phonemes and features have an extremely strong effect on our simulated acoustic and articulatory measurements. Further experimental and modelling investigation of this finding is required.

Finally, chapter 9 summarises these findings and suggests potential methodological and theoretical developments for the future.

Chapter 2

Investigating and extending Dell's (1986) model of speech errors

2.1 Introduction

One of the most influential models of the production of speech and of speech errors is Dell's (1986) spreading activation model. The present thesis examines and extends the model in the face of existing and new evidence.

The first section of this literature review summarises some key findings from investigations rooted in the transcribed speech error literature. We look at theories of how sounds move from one position to another, and arguments about information flow between the lexical selection and phonological encoding stages of the word production system. It is concluded that Dell's (1986) model is the most comprehensive model in this area.

The second section reviews the assumptions made in Dell's (1986) model about output from phonological encoding and the possibility of errors at a subphonemic level, and highlights problems with basing these assumptions on perceptual data. The section then outlines some new acoustic and articulatory experiments which address the question of how information flows between phonological encoding and subphonemic processes. Some alternative explanations of this new data are proposed, which are suitable for investigation by simulation.

The third and final section of this literature review asks how we should go about investigating information flow options in a model with so many free parameters. It reviews the literature on the effects of manipulating parameters in Dell's (1986) spreading activation model, and evaluates previous approaches to comparing architectural options in this model.

2.2 Speech error evidence and models: lexical selection and phonological encoding

How do we produce words? In normal conversation, we generate words with ease, uttering around 150 per minute (Maclay & Osgood, 1959). But around once every 1000 words, the process goes audibly wrong (Garnham, Shillcock, Brown, Mill, & Cutler, 1981). By examining the form of these speech errors and specifying what makes them more likely to occur, we can shine some light on the representations and processes used in the human word production system.

Speech errors are usually characterised as unintended components intruding on or replacing part of the intended production. Different sized components of the utterance can be erroneously encoded (e.g., Fromkin, 1971; Garrett, 1975). For instance, example 1a demonstrates the exchange of two words, and example 1b demonstrates the exchange of two phonemes. Unintended components often have a clear origin nearby in the utterance, with analyses of phoneme errors suggesting that between 75% to 95% of errors involve units from elsewhere in the speech plan (del Viso, Igoa, & Garcia-Albea, 1991; Pérez, Santiago, Palma, & O'Seaghdha, 2007; Shattuck-Hufnagel & Klatt, 1979; Vousden, Brown, & Harley, 2000).

(1a) *"Although suicide is a form of murder"* → *"Although murder is a form of suicide"* (Garrett, 1975)

(1b) *"cold hard cash"* → *"hold card cash"* (Fromkin, 1973)

These results imply that production of a phrase or word involves a number of processes (e.g. Dell, 1986; Fromkin, 1971; Garrett, 1975; Shattuck-Hufnagel, 1979), including selection of the intended word, a process known as *lexical selection*, and selection of the sounds required for that word, a process known as *phonological encoding*. The output of each of these processes is assumed to be an ordered string of units of the appropriate size; words for lexical selection, and phonemes for phonological encoding. If the process malfunctions, units may replace each other, resulting in speech errors such as those reported in examples 1a and 1b.

In this section, we examine empirical evidence and theoretical accounts of this evidence in two steps. The first part of this section focuses on accounts of how units are ordered in word production, and explanations of the frequently observed misorderings, or movement errors. In the second part, the interaction of the lexical selection and phonological encoding processes is considered in more depth.

2.2.1 Accounts of movement errors

The earliest account of movement errors was the frame and slot model, as outlined in detail by Shattuck-Hufnagel (1979), and ideas from this model continue to dominate later accounts. The most influential development of this theory was Dell's (1986) spreading activation account. This section outlines the mechanics underlying Shattuck-Hufnagel's (1979) and Dell's (1986) accounts of movement errors, and evaluates some core evidence which has been provided in support of them, before providing a brief summary of other proposed models of unit misordering.

Shattuck-Hufnagel's (1979) frame and slot model

In Shattuck-Hufnagel's (1979) model, processing at each level operates upon a buffer, which contains units required for the output of previously prepared higher level units. For example, at phonological encoding, the buffer would contain phonemes needed to encode words which have already been selected by the lexical selection process. The use of a buffer is motivated by the occurrence of errors in which units are produced too early. These errors suggest that higher level processes are further progressed in preparing the utterance than lower level processes, and that the higher level processes make units available to lower level processes earlier than the point at which they are to be produced (e.g., Fromkin, 1971; Lashley, 1951; MacKay, 1970; see also Dell, Burger, & Svec, 1997 for a recent discussion of speech errors and buffers). An ordering mechanism then selects units from the buffer for output in the correct order.

Movement errors can be divided into three categories: *anticipation* of an upcoming unit (as in example 2a), *perseveration* of a previous unit (as in example 2b), and the full *exchange* of two units (as in example 2c).

(2a) *"take my bike"* → *"bake my bike"* (Fromkin, 1973)
(2b) *"gave the boy"* → *"gave the goy"* (Fromkin, 1973)
(2c) *"copy of my paper"* → *"poppy of my caper"* (Fromkin, 1973)

Some evidence suggests that exchange errors are less frequent than anticipations and perseverations (e.g., Nooteboom, 1969; Stemberger, 1989). However, data from phoneme substitutions shows that even by the most conservative analyses, exchange errors comprise at least 5% of all movement errors (Stemberger, 1989). An account of exchange errors which assumes that exchange errors are simply the serendipitous occurrence of two symmetrical substitutions would therefore not appear convincing

(Shattuck-Hufnagel, 1979), especially given the overall low occurrence rate of speech errors (one or two every thousand words; Garnham et al., 1981). A model of ordering is required in which the occurrence of an anticipation increases the chance of the complementary substitution.

The frame and slot ordering mechanism described by Shattuck-Hufnagel (1979) meets this specification. In the frame and slot model, frames describe the necessary shape of the ordered output, such as a phrase, or a syllable. Slots in these frames, and the units in the buffer which fill these slots, are marked with content attributes. At lexical selection, the syntactic class of the slots and fillers is indicated, and for phonological encoding, slots and fillers are marked with their syllable position.

Observing the content markup, the *scan-copier* mechanism selects a unit from the buffer appropriate for each slot in the frame. Once a unit has been allocated to a slot, it is marked as used by the *checkoff monitor* and is no longer available for selection by the scan-copier. For example, to phonologically encode the phrase *"big fun"*, the scan-copier will locate the phoneme /b/, allocate it to the onset slot of the first word, and then the checkoff monitor will mark the phoneme as produced. The scan-copier will then proceed to select the phoneme /ɪ/, and so on. The marking of slots and fillers for syllable position or syntactic class allows the model to capture the result that words are generally replaced by words of the same class (e.g., Fay & Cutler, 1977), and that misordered phonemes tend to change word but not syllable position (e.g. Shattuck-Hufnagel, 1979).

Exchange errors are easily explained within this framework. If during production of the phrase *"big fun"*, the scan-copier misselects /f/ and places it in the onset slot for the first word, and the checkoff monitor marks the phoneme as produced, then /f/ will not be available for production at the onset of the second word. As the phoneme /b/, marked as an onset, will be available in the buffer of phonemes, the scan-copier may be forced to resort to this phoneme instead, thereby triggering the second substitution and producing the exchange error *"fig bun"*.

Anticipations and perseverations can also be explained within this framework. Anticipations and exchanges begin with the same error, where in the *"big fun"* example, the scan-copier misselects /f/ and places it in the onset slot for the first word. In an anticipation however, the checkoff monitor does not mark /f/ as produced. The /f/ phoneme is then still available for the onset slot of the second word, and its selection results in the anticipation *"fig fun"*. A perseveration occurs when the scan-copier correctly selects /b/ for the onset slot of the first word, but the checkoff monitor

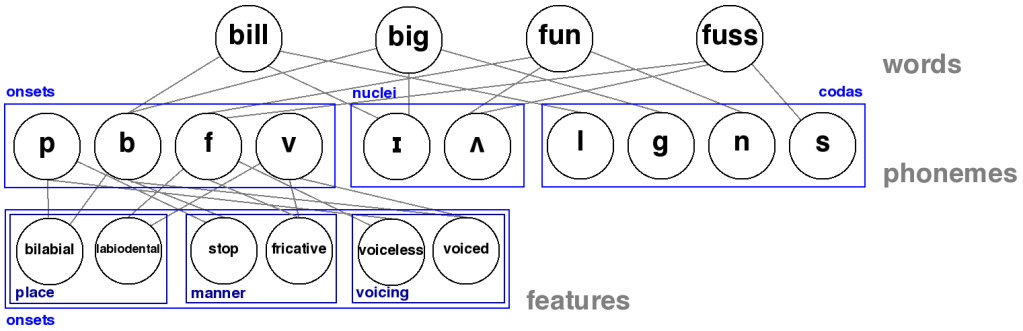


Figure 2.1: An excerpt of the model proposed by Dell (1986), displaying some units involved in phonologically encoding the words *"big"* and *"fun"*. No nucleus and coda features are shown.

does not check /b/ off (or delays the checkoff), so that /b/ is still available to be selected at the onset slot of the second word. If the scan-copier then misselects /b/ for this position, then the perseveration error *"big bun"* will occur.

Dell's (1986) spreading activation model of phonological encoding

Dell (1986) built on Shattuck-Hufnagel's (1979) work by providing an implementation of the frame and slot model, using *spreading activation* (e.g., McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982). Whilst the theory presented by Dell (1986) spanned the entire sentence production process, the implemented model focused on phonological encoding. Later implementations have incorporated lexical selection as well as phonological encoding (e.g., Dell, Schwartz, Martin, Saffran, & Gagnon, 1997), although these have focused on single word production. In Dell's (1986) model, phoneme nodes are split into onset phoneme nodes, nucleus phoneme nodes, and coda phoneme nodes. One phoneme is chosen from each of these groups to fill the onset, nucleus and coda slots in a theoretical CVC syllable frame (see Dell, 1988 and Hartsuiker, 2002 for a variation upon this model which permits syllables of varying shapes to be produced). Activation of phoneme nodes represents their presence in the phoneme buffer, and the scan-copier mechanism is implemented by a process which selects the most activated node. To complete the model, the check-off monitor functionality is provided by a post-production inhibition system which reduces the activation of a phoneme to zero immediately following production.

An excerpt of the model proposed by Dell (1986) is depicted in figure 2.1, showing some of the units involved in phonologically encoding the phrase *"big fun"*. Production of this phrase begins with the word node *big* being marked as selected, and receiving a large amount of activation. The *big* node passes activation to its component phonemes, including the onset phoneme /b/. The upcoming word *fun* is also *primed* with a small amount of activation, reflecting that this word has been planned for production by lexical selection. The *fun* node therefore also conveys a small amount of activation to its component phonemes. Next, phoneme selection takes place. In a correct production, /b/ will be the most active onset phoneme, and will therefore be selected for production. Following selection of the onset, nucleus and coda phonemes, the activation level of the produced representation are set to zero, completing production of this word. The next word is produced in a similar manner, beginning with selection and full activation of the word node *fun*. Activation again passes from the word layer to the phoneme layer. Phoneme selection then proceeds once more, and in a correct production, /f/ will be the most activated onset phoneme and selected for production as the onset of the second word. Selected representations have their activation levels set to zero, and production of the phrase *"big fun"* is complete. Again, transposed phonemes generally maintain their syllable position. As phonemes in different positions are represented separately, the activation of a phoneme unit in one slot cannot affect the activation of a unit representing the same phoneme in another slot.

Exchanges, anticipations and perseverations are explained by relying on the concepts of selection and check-off, or post selection suppression, reflecting the frame and slot model origins of the theory. In an exchange, the priming of the /f/ node, coupled with noise in the network, causes the /f/ node to receive more activation than the intended /b/ node. This leads to the misselection of the /f/ node. Selection in turn results in the suppression of the activation levels of /f/. The intended phoneme /b/ is not suppressed, however, as it was not selected. The following word *fun* is then activated, and activation passed to its component phonemes. However, in an exchange, the activation retained by the unselected /b/ node, again aided by noise in the network, means that the activation passed to /f/ from *fun* is not enough to secure the position of /f/ as the most activated node. Instead, /b/ is selected as the onset of the second word, resulting in the exchange error *"fig bun"*.

The mechanism used to generate anticipations overlaps considerably with the mechanism used to generate exchanges. Misselection of the /f/ node occurs in the same way as in an exchange, and /f/ is again suppressed whereas /b/ is not. During

production of the second word however, different random noise conditions in the network mean that the activation retained by /b/ is not sufficient to outweigh the activation conveyed from *fun* to the phoneme /f/. The /f/ phoneme is therefore selected as the onset of the second word, producing the anticipation *"fig fun"*.

A perseveration occurs in a rather different way to an exchange. In a perseveration, /b/ is correctly selected for output in the first onset position, and its activation set to zero following production. However, other words which begin with /b/ such as *bill* and *bat* will have acquired activation during production of the word *big*, via feedback connections from the onset node /b/. In a perseveration, activation from these neighbouring words to the onset node /b/, alongside suitable noise conditions, leads /b/ to acquire more activation than the intended phoneme /f/, thereby resulting in the error *"big bun"*.

Support for Dell's (1986) account of movement errors

A core part of the support of Dell's (1986) explanation of movement errors stems from the similarity between the model's behaviour and speech corpora results reported by Nooteboom (1969). In Nooteboom's (1969) data, anticipations are vastly more common than perseverations, and perseverations in turn far outnumber exchanges. Similarly, in Dell's (1986) model, anticipations are the most frequent kind of error, as error generation depends only on noise in the network and the priming which is given to all upcoming phonemes. Perseverations occur less frequently than anticipations, as the average amount of activation conveyed to a previously produced phoneme by the intended first word's neighbours is less than the amount of priming activation provided to an upcoming word. Finally, exchanges are the least frequent error of all, as noise conditions must be sufficient to both cause an error on the first word, and then allow this error to successfully trigger a further error on the second word.

However, there is an important observation to be made here. This behaviour of Dell's (1986) model, in which exchanges are by far the least frequent error, underlines a key difference between this model and Shattuck-Hufnagel's (1979) upon which it was based. In Shattuck-Hufnagel's (1979) model, only one error in the scan-copier is required for an exchange to be generated. The normal behaviour of the checkoff mechanism in marking the anticipated phoneme as produced will then lead to an exchange. In contrast, an anticipation or perseveration requires two errors; both a misselection by the scan-copier, and a checkoff error by the checkoff mechanism, allowing a produced phoneme to be produced again. Hence,

Table 2.1: Proportions of movement errors reported in Nooteboom (1969) and Shattuck-Hufnagel and Klatt (1979)

	Anticipations	Perseverations	Exchanges	Incompletes
Nooteboom (1969)	76%	17%	7%	–
Shattuck-Hufnagel and Klatt (1979)	10%	19%	24%	47%

Shattuck-Hufnagel's (1979) model predicts that exchange errors would occur most frequently.

This crucial difference in the behaviour of the models in turn reflects a disparity in the different sets of data and interpretation of the data around which the models were developed. Shattuck-Hufnagel (1979) took the constraints for her model from the MIT-CU corpus, reported on in Shattuck-Hufnagel and Klatt (1979), whereas Dell (1986) compared his model's behaviour to data reported by Nooteboom (1969). The two sets of data are shown in table 2.1. In the table, counts of anticipations, perseverations, exchanges and *incomplete errors* are shown. Incomplete errors are errors such as *"big fun"* → *"fig... big fun"*, where the speaker anticipates a phoneme but then stops and corrects themselves rather than continuing with the utterance. As the data reported by Shattuck-Hufnagel and Klatt (1979) suggest, these form a sizeable portion of human phonemic speech errors. Nooteboom (1969), in this early corpus, does not provide a separate count of incomplete errors.

Table 2.1 shows that, in proportional terms, over three times as many exchanges were found in Shattuck-Hufnagel and Klatt (1979) than in Nooteboom (1969), such that Shattuck-Hufnagel and Klatt (1979) found more exchanges than perseverations, unlike Nooteboom (1969). However, a further difference between the data is the interpretation of incomplete errors. Without the completion of an incomplete error, it is not intrinsically clear whether such errors represent an incomplete anticipation (*"big fun"* → *"fig fun"*) or an incomplete exchange (*"big fun"* → *"fig bun"*) (e.g Cutler, 1981; Dell, 1986; Dell & Reich, 1981; Fromkin, 1971; Garrett, 1975; Nooteboom, 1980, 2005b; Shattuck-Hufnagel, 1979; Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1989). Shattuck-Hufnagel (1979), in her interpretation of the data, argues that these errors are exchanges, noting that in the MIT-CU corpus, the target and error phonemes in anticipations and perseverations nearly always share a large number of features, whereas this constraint is weaker for both exchanges and incomplete errors. Adding the number of incomplete errors found to the exchange category would clearly make exchanges by far the most common kind of movement error. However, the figures rather suggest that Nooteboom (1969) may have chosen the alternative interpretation and classified all incomplete errors

as anticipations, leaving many more anticipations than perseverations or exchanges. As such, it seems likely that the classification of incomplete errors is playing a very important role in determining the reported patterns of data, which these models are in turn compared to. The questions of how this data should actually be interpreted and possible further inconsistencies between data from different corpora are returned to in chapter 5, where the relationship of Dell's (1986) model to the empirical evidence is reexamined.

Other accounts of movement errors

Very few other accounts of movement errors have been proposed. A large amount of the speech error modelling literature consists of extensions of Dell's (1986) model which produce one word only (e.g., Dell, Schwartz, et al., 1997; Foygel & Dell, 2000; Rapp & Goldrick, 2000). Hartsuiker (2002) presents an implemented version of Dell (1986, 1988) which produces more than one word and successfully generates movement errors, but this model does not aim to extend the theory of movement error generation. Various other spreading activation models in a similar vein to Dell (1986) have been suggested, but most of these are either not implemented (e.g., MacKay, 1987; Stemberger, 1985), produce only one word (e.g., Harley, 1993), or do not report in detail on exchange errors and other movement errors (e.g., Schade & Berg, 1992).

Only two other implemented models attempt to simulate anticipations, perseverations, and exchanges. Firstly, Vousden et al. (2000) provide an implemented dynamic oscillator account. There are big differences between the processing mechanics of Dell's (1986) spreading activation model and the dynamic oscillator model, yet both models include the concept of selection of the most activated representations, and suppression of recently produced representations. Furthermore, in both cases, the suppression of an anticipated phoneme and the lack of suppression of the intended but not produced phoneme provides a trigger for exchange errors. Vousden et al.'s (2000) model also replicates the qualitative pattern observed by Nooteboom (1969).

Secondly, Levelt, Roelofs, and Meyer (1999) present a simulation of movement errors within the WEAVER++ model. This model in its natural state produces no errors at all, due to a verification procedure carried out by syllable nodes, which in this model are situated between phonemes and articulatory processes. In production of the phrase *"big fun"*, the component phonemes of the word *"big"* would receive activation, and the verification procedure on the articulatory syllable node for *big*

would only pass if activation from an onset /b/, a nucleus /ɪ/ and a coda /g/ was received, where all these phonemes were marked as participating in the current syllable. Levelt et al. (1999) suggest that errors may result from this verification procedure failing. For example, the anticipation error *"big fun"* → *"fig fun"* may be produced if the verification procedure for the *fig* syllable fails to notice that the /f/ is from a different syllable. A similar explanation exists for the perseveration error *"big fun"* → *"fig fun"*.

A core weakness of this explanation however, as highlighted by Levelt et al. (1999) and their simulation data, is its tendency to generate almost no exchange errors. Examination of the theory proposed suggests that there is no trigger mechanism, such that the anticipation error *"big fun"* → *"fig"* does not render the completion *"bun"* more likely. Earlier in this section, it was highlighted that a model of exchange errors attributing the occurrence of the two errors to coincidental symmetry was unlikely to be correct, and this model fails for exactly this reason. With reference to their simulation data, Levelt et al. (1999) do however claim that the model generates more anticipations than perseverations, in line with Nooteboom's (1969) results. However, no clear explanation of the theoretical reasons for this anticipatory bias is presented. Presumably it is not the case that the probability of a verification error has been set to a higher value for the first syllable than for the second. There is little obvious motivation for such a feature of the word production system, nor is it at all evident that such a simplistic approach would scale to account for the relevant proportions of movement errors if longer productions were attempted.

Anticipations, perseverations, and error rate

Finally, we note Dell, Burger, and Svec's (1997) abstract investigations into models of language production, and the relationship between generation of anticipation and perseveration errors and overall error rate. In this research, the model proposed is intentionally extremely abstract and the generation of exchange errors is not considered, but the empirical and theoretical results of this study make useful contributions to the study of models of movement errors which we return to later in this thesis.

Dell, Burger, and Svec (1997) highlight that across a variety of results, higher error rates are accompanied by lower proportions of the errors being anticipations rather than perseverations. Specifically, aphasic patients (Schwartz, Saffran, Bloch, & Dell, 1994), children (Stemberger, 1989), and people who are under pressure to speak

quickly (Dell, 1990; Dell, Burger, & Svec, 1997) exhibit higher error rates, and lower proportions of anticipatory errors. Conversely, practising production of a specific phrase leads to lower error rates, and higher proportions of anticipatory errors (Dell, Burger, & Svec, 1997; Schwartz et al., 1994). Schwartz et al. (1994) further observe that speakers who exhibit "good" behaviour patterns, i.e. lower error rates and higher proportions of anticipatory errors, also produce higher proportions of word outcome errors rather than non-word outcome errors.

Dell, Burger, and Svec (1997) present an abstract model of language production, in which a node representing either the past, the present or the future is selected for output. Dell, Burger, and Svec (1997) argue that this abstract framework allows them to capture the basic important behaviour of any successful model of sequence production, whereby the model must activate the representation to be produced in the present, deactivate representations produced in the past, and prepare to activate representations to be produced in the future. The mechanism outlined by Dell (1986) clearly fits in with this abstract description, by jolting representations to be produced immediately, inhibiting representations which have been produced in the past, and priming representations to be produced in the future. Dell, Burger, and Svec (1997) show that in this abstract model, it generally holds that parameter settings which make the model generate more errors also cause the model to produce a lower proportion of anticipations. The effect of manipulating the parameters can be determined by rearranging equations which completely describe the behaviour of this very abstract and simple model.

The parameters of the model include: the strength of the connections from the abstract representations used for planning the current output to the less abstract representations which are selected for output (where connection strength is assumed to represent how well a sequence has been learnt), the number of timesteps at which the activation of the representations is calculated before the output of the model is determined (where lower numbers of steps are assumed to represent faster speech rates), the rate at which activation of a representation decays, the amount of activation with which representations to be produced in the future are primed, and parameters which govern how noisy the decision process is. The equations of the model demonstrate that the model's tendency to generate anticipations is determined only by the amount of activation passed to future representations, and how noisy the decision process is. Manipulations of other parameters only affect the probability that the model will produce a perseveration. Therefore, if the amount of activation priming and the parameters governing the decision process remain

constant, both the error rate and the proportion of errors which are anticipatory are determined by the number of perseverations generated by the model, so that there is a negative correlation between error rate and the proportion of anticipatory errors as in the empirical results. For example, Dell, Burger, and Svec (1997) show that increasing the strength of the connections from abstract to less abstract representations, which they assume results from practising the sequence for production, both reduces the error rate and increases the proportion of errors which are anticipatory, mimicking the empirical effect of practice reported by both Schwartz et al. (1994) and Dell, Burger, and Svec (1997). Increasing the number of steps which pass before output occurs, where a higher number of steps is assumed to correspond to a slower speech rate, also reduces the error rate and increases the proportion of errors which are anticipatory, again fitting in with Dell's (1990) and Dell, Burger, and Svec's (1997) empirical results. Increasing the rate at which the activation of representations decays reduces the influence of the past on the present, thereby also reducing the tendency of the model to generate perseverations. At higher decay rates, the error rate is therefore lower and the proportion of anticipatory errors generated higher, again in line with the variation seen in the empirical results.

Whilst this model is extremely abstract and does not offer an account of exchange errors, we will return to this empirical result and the theoretical insight gained into desired model behaviour in chapter 5.

2.2.2 Information flow between lexical selection and phonological encoding

The introduction to this section provided examples of both word and phoneme misorderings. To explain these two types of errors, it has been suggested that a two stage lexical selection and a phonological encoding model is necessary (e.g., Garrett, 1975; Nooteboom, 1969). Of the models examined in section 2.2.1, both the Vousden et al. (2000) model and Dell's (1986) original implementation cover phonological encoding only. However, others have covered both stages, including the model described by Levelt et al. (1999), and extensions of Dell's (1986) model which encompass more of the theory outlined in Dell's (1986) paper; in particular, Dell, Schwartz, et al. (1997).

These latter models are both spreading activation models, in which semantic features pass activation to word representations at the lexical selection stage, and these lexical representations in turn activate phoneme representations at the phonological encoding stage. However, an important difference exists between these models,

concerning the way in which activation passes between representations at different stages.

In Levelt et al.'s (1999) model, the lexical selection and phonological encoding stages are independent, or *discrete*. Goldrick (2006) defines discrete systems as obeying three constraints. Firstly, processing at a given stage (e.g., phonological encoding) only begins once selection has occurred at the previous stage (e.g., lexical selection). Secondly, only selected representations pass activation on to subsequent stages. Thirdly, activation flows from one stage to subsequent stages, but does not flow back.

In contrast, the model presented by Dell, Schwartz, et al. (1997) can be described as *interactive*. Detailed aspects of processing at lexical selection affect phonological encoding, and vice versa. Specifically, activation *cascades* from lexical selection to phonological encoding, and *feeds back* from phonological encoding to lexical selection. A cascading system removes the first two constraints of a discrete system (Goldrick, 2006). Before selection occurs at a given stage (such as lexical selection), a cascading system will permit activation to flow to subsequent stages (such as phonological encoding), whereas a non-cascading system will not. After selection has occurred, a cascading system will permit activation to be conveyed from non-selected representations, whereas a non-cascading system will not. In this thesis, we are particularly concerned with the latter aspect of cascading, that activation can be transmitted from representations which have not been selected for production. Feedback systems assume cascading, and further remove the third constraint of a discrete system, such that activation flows back from lower level stages (such as phonological encoding) to higher level stages (such as lexical selection).

Goldrick (2006) has argued for interaction between lexical selection and phonological encoding as hypothesised by Dell, Schwartz, et al. (1997), with the proviso that this interaction must be limited (e.g., Goldrick, 2006; Rapp & Goldrick, 2000). We begin by considering arguments for cascading from lexical selection to phonological encoding, move on to arguments for feedback from phonological encoding to lexical selection, and conclude with a note about monitor based explanations of the outlined evidence.

Cascading from lexical selection to phonological encoding

A core piece of evidence in the argument for cascading from words to phonemes is the rate at which speakers generate mixed errors, such as *"cat"* → *"rat"*. Chance would

predict that formal errors (that is, errors which differ minimally in phonological form from the target) should also share semantic features with the target as often as phonological neighbours of a target are also semantic neighbours. However, corpus analyses (Dell & Reich, 1981; Harley, 1984) and experimental investigations on both normal speakers (Ferreira & Griffin, 2003; Martin, Weisberg, & Saffran, 1989) and aphasics (Martin, Gagnon, Schwartz, Dell, & Saffran, 1996; Rapp & Goldrick, 2000) have shown that speakers generate many more mixed errors than this.

A discrete model struggles to account for this evidence, whereas the result falls out naturally from a model in which activation cascades from words to phonemes (e.g., Rapp & Goldrick, 2000; Goldrick & Rapp, 2002; Goldrick, 2006). In both models, to produce the word *"cat"*, semantic features such as *furry*, *feline*, and *pet* may be activated. This combination of features will render the word *cat* the most active at the lexical level. However, the word *rat* will also receive some activation from the semantic feature *furry*. Activation from the *cat* node then passes on to its component phonemes /k/, /æ/ and /t/.

Crucially, in a model in which activation cascades from the lexical level to the phonological level, the phoneme /r/ will receive activation from the activated semantic neighbour *rat*. A cascading model predicts that the node /h/ will not receive such support from *hat*, as *hat* is not semantically related to *cat*. Conversely, in a discrete model, neither /r/ nor /h/ will receive any activation from the lexical level, as they are not selected.

In this way, a cascading model predicts that phonemes which form mixed errors such as *rat* are more likely to be erroneously selected at the phoneme level than phonemes which form purely formal errors such as *hat*, whereas a discrete model does not. The ability of cascading models to generate this pattern of behaviour has been verified by simulations (Rapp & Goldrick, 2000).

More support for this claim is provided by evidence demonstrating that formal errors result in words of the same syntactic category more often than chance would predict (e.g., del Viso et al., 1991; Dell, Schwartz, et al., 1997; Fay & Cutler, 1977; Gagnon, Schwartz, Martin, Dell, & Saffran, 1997) This is easily explained in a similar manner, if it is assumed that words also receive activation from syntactic features to ensure that a syntactically appropriate word is selected (e.g., Goldrick & Rapp, 2002). In a cascading model, production of the word *"cat"* would involve activation passing from a *noun* feature to all nouns. If activation cascades from the lexical level to

the phonological level, then this extra activation will be conveyed to the phonemes in the noun /h æ t/, but not to the phonemes in the verb /s æ t/. Simulations presented by Goldrick and Rapp (2002) again confirm that this explanation is able to account for the data.

However, simulation data reported by Goldrick (2006) indicates that cascading from the lexical level to the phoneme level must be limited. Goldrick (2006) defines the strength of cascading as inversely proportional to *selection strength*. In a system with strong selection strength, or limited cascading, the activation passed from selected items to subsequent levels will be much stronger than the activation conveyed from unselected items. Conversely, a system with weak selection strength, or strong cascading, will display less differentiation between the levels of activation transmitted from selected and unselected items.

Goldrick's (2006) simulations demonstrate that in models with lexical level damage, strong cascading leads to a very high percentage of nonword productions, because the activation levels of the phonemes required to produce the word are not boosted sufficiently in comparison to other phonemes. This contradicts evidence from aphasia investigations, which show that patients with deficits localised to the lexical level do not generate nonword errors (e.g., Goldrick & Rapp, 2002; Rapp & Goldrick, 2000). Models with more limited cascading do not demonstrate this behaviour, thereby providing a better fit to the data. Goldrick (2006) further highlights that the hypothesis of limited cascading is in line with chronometric results (e.g., Peterson & Savoy, 1998; Levelt, Schriefers, Vorberg, Meyer, & Pechmann, 1991), where evidence of cascading is only found when there is very strong semantic similarity for semantic neighbours, very high phonological similarity for phonological neighbours, or simultaneous semantic and phonological overlap.

Feedback from phonological encoding to lexical selection

However, there is some evidence which does not naturally fall out of a purely feedforward cascading model. The most prominent example is the *lexical bias* effect. The lexical bias effect refers to the observation that errors are more likely to result in real word outcomes than chance would predict. This effect has been demonstrated by both experimental investigations and corpus analyses (e.g., Baars, Motley, & MacKay, 1975; Dell, 1986; Dell & Reich, 1981; Hartsuiker, Corley, & Martensen, 2005; Humphreys, 2002; Nooteboom, 2005a, 2005b; though see del Viso et al., 1991 for a null result). A feedback model easily explains this result.

Consider production of the word *"cat"*. In both discrete and interactive models, following lexical selection, activation will pass from the *cat* node to the phonemes /k/, /æ/ and /t/. In a feedback model, however, activation will spread back from these phonemes to other words which these phonemes participate in, such as *sat*. Activated words will then activate their component phonemes, such as /s/. In a feedback model, phonemes in a word like *sat* will therefore be more activated than phonemes in a non-word such as *"lat"*, which by definition has no representation at the lexical level. This difference will lead to a lexical bias in errors. Conversely, in a discrete model, this difference in activation will not exist. The ability of feedback models to account for lexical bias evidence has been demonstrated using simulations (e.g., Dell, 1986; Rapp & Goldrick, 2000).

Further support for feedback from phonological encoding to lexical selection is provided by evidence of a mixed error effect in aphasic patients with damage localised to the lexical level (Rapp & Goldrick, 2000; Goldrick, 2006). If errors are known to be occurring at the lexical level rather than the phonological level, then the explanation of the mixed error effect given in the previous section, which relied on cascading from semantic competitors at lexical selection to their phonological representations at the phonological encoding stage, cannot be enough. However, increased misselection of mixed competitors at the lexical level (e.g. *"cat"* → *"rat"*) in comparison to purely semantic competitors (e.g. *"cat"* → *"dog"*) can easily be accounted for by feedback from phonemes in the target word (e.g. /æ/ and /t/) to lexical selection (e.g. Dell, Schwartz, et al., 1997). Simulations reported by Rapp and Goldrick (2000) confirm that the pattern of errors generated by such aphasic patients can only be accounted for when feedback from the phonological level is present.

It has also been shown that words which have more phonological neighbours are less susceptible to errors and are produced more quickly (e.g., Vitevitch, 2002). Broadly speaking, phonological neighbours are words which have only a minimal phonological difference from the target; for example, *"can"* and *"rat"* are both neighbours of *"cat"*. Vitevitch (2002) showed that these results still hold when sublexical properties such as phoneme identity and phonotactic probability are controlled for, suggesting that this effect must arise at the lexical level. For phonological characteristics of the words to affect lexical selection, feedback from phonological encoding to the lexical level is again necessary. In this model, feedback reverberations from *can* and *rat* via the phonemes they share with *cat* lead to increased activation of *cat*, which in chronometric models (e.g. Levelt et al., 1999) would increase the

speed with which the representation was selected, and in models of speech errors, would decrease the number of errors in encoding (e.g. Dell, 1986; Dell, Schwartz, et al., 1997). Simulations reported by Dell and Gordon (2003) appear to confirm the prediction of the speech error model.

Like cascading, however, feedback must be limited. Simulations reported by Goldrick (2006) demonstrate that in a model with lexical level damage, models with strong feedback from phonological encoding to lexical selection generate a large number of formal errors, such as *"cat"* → *"hat"* (in Goldrick's (2006) simulations, particularly models which have much stronger feedback than feedforward connectivity). This is not the result found in aphasic patients with damage localised to the lexical level, who do not produce any phonological errors (Rapp & Goldrick, 2000; Goldrick, 2006). A model in which feedback from the phoneme level is not strong enough to give phonological neighbours a noticeable advantage over non-neighbours for selection at the damaged lexical level does not suffer from this problem (Goldrick, 2006).

Monitoring

The previous two sections have outlined arguments from the literature for interactivity between lexical selection and phonological encoding. It has however been proposed that a discrete model could account for a number of these effects by employing a monitor and editor, usually relying on representations and processes already employed by the comprehension system (e.g. Baars et al., 1975; Levelt, 1983, 1989; Levelt et al., 1999; Nooteboom, 2005a, 2005b). For example, lexical bias could be explained by a monitor recognising and editing out more non-word outcome errors than word outcome errors. The mixed error effect could be explained by arguing that the monitor would be more likely to miss errors which were both semantically and phonologically similar to the target than errors which were either just semantically similar to the target or just phonologically similar to the target.

Despite the age of this suggestion, to date no implementation exists which accounts for the lexical bias or the mixed error effect. This reflects the current vagueness of the proposal, which leaves it somewhat overpowerful and difficult to falsify (see similar criticisms in e.g., Goldrick & Rapp, 2002; Rapp & Goldrick, 2000). The aim here is not to argue against the role of a monitor in word production. It appears clear that even basic speech error results, such as the existence of incomplete errors, require a monitor and editor for a full explanation to be achieved (and this point is revisited in chapter 5). However, pending more detailed specification of

this mechanism, this thesis takes the approach that it would not be productive to simply accept this unimplemented proposition and in particular, to allow it to rule out other more clearly described explanations, preventing further investigation of their predictions. Indeed, lexical bias results reported by Hartsuiker et al. (2005) suggest that a complete model of word production would involve a mixture of both interactive processing and editing mechanisms.

2.2.3 Summary

This section summarised some key findings from empirical and modelling investigations of speech errors. It began by noting that speech errors can often be described as misorderings of the speech plan (e.g., del Viso et al., 1991; Pérez et al., 2007; Shattuck-Hufnagel & Klatt, 1979; Vousden et al., 2000). Furthermore, different sized units can be involved in speech errors, such as words and phonemes, suggesting that a number of processes are involved in word production, including lexical selection and phonological encoding (e.g., Dell, 1986; Fromkin, 1971; Garrett, 1975; Shattuck-Hufnagel, 1979).

The first part of this section looked at theories of how misorderings of the speech plan occur. It was noted that more exchange errors (e.g., *"big fun"* → *"fig bun"*) occur than could be explained by the coincidental occurrence of two symmetrical substitutions (Shattuck-Hufnagel, 1979). In a successful account, an anticipation (e.g., *"big fun"* → *"fig. . . "*) should therefore make the substitution required for a complete exchange (i.e., *"bun"*) more likely than it would be if the anticipation had not occurred. The frame and slot theory described by Shattuck-Hufnagel (1979) meets this constraint. However, very few implemented models account for movement errors. One implementation fails to generate any exchanges due to the constraint noted above not being met, and anticipations not triggering the completion error (Levelt et al., 1999). The two which do successfully generate exchanges are both implementations of the frame and slot theory (Dell, 1986; Vousden et al., 2000). Both models generate the same pattern of relative proportions of anticipations, perseverations and exchanges as found by Nooteboom (1969), adding validity to the explanations proposed. However, we note that there may be some problems with the selection and interpretation of Nooteboom's (1969) data as an empirical reference point, particularly as other corpora suggest other patterns (e.g., Shattuck-Hufnagel & Klatt, 1979), and it is not clear if the classification of incomplete errors such as *"big fun"* → *"fig. . . big fun"* was sensible. Chapter 5 reexamines both the interpretation of data from multiple corpora and the relationship of Dell's (1986)

model to the empirical evidence, and also investigates to what extent the model exhibits the relationship between generation of anticipation and perseveration errors and error rate as reported in Dell, Burger, and Svec's (1997) empirical and abstract modelling investigations.

The second part of this section considered models which can account for both word and phoneme errors by positing a lexical selection and a phonological encoding stage, in particular the spreading activation models suggested by Levelt et al. (1999) and Dell, Schwartz, et al. (1997), where the latter is an extension of Dell's (1986) model. Specifically, the question of information flow between these two processes was briefly examined. Concepts of discrete and interactive systems, including the ideas of cascading and feedback, were outlined. A number of error patterns and other evidence was summarised, and it was concluded that these patterns can be accounted for if cascading from lexical selection to phonological encoding, and feedback from phonological encoding to lexical selection is assumed, as implemented in Dell, Schwartz, et al.'s (1997) model. The likely role of monitoring (e.g., Levelt, 1983, 1989; Levelt et al., 1999) in word production was acknowledged, but no implementation of the monitor has been provided which can explain all the effects which interactive processing simulations account for.

Dell's (1986) original model focused solely on phonological encoding, and extensions of Dell's (1986) model encompassing lexical selection have concentrated on single word production (e.g., Dell, Schwartz, et al., 1997; Foygel & Dell, 2000; Goldrick, 2006; Rapp & Goldrick, 2000). There is therefore no implementation of Dell's (1986) theory which simultaneously captures movement errors and results such as the mixed error effect, which rely on interactivity between the lexical selection and phonological encoding processes. Despite the divided nature of these simulations, Dell's (1986) theory as a whole still remains by far the most successful at explaining and simulating such a wide range of speech error results, and this success has made it the most influential model of word production in the speech error tradition. For the same reasons, this model forms the focus of the rest of this thesis.

2.3 Beyond the phoneme

The previous section considered the lexical selection and phonological encoding processes, and information flow between the two. But how does information flow from phonological encoding to subphonemic processes?

The model presented by Dell (1986; Dell, Schwartz, et al., 1997) posits that the the output of phonological encoding is a string of phonemes. Dell's (1986) original model does include a layer of features below the phoneme layer. However, selection does not occur at this layer, and output is read from the phoneme layer. The presence of features in the model is motivated by the *phonological similarity effect*, the result that more similar phonemes are more likely to exchange (e.g., Levitt & Healy, 1985; MacKay, 1970; Nooteboom, 1969). If feedback from features to phonemes is assumed, a target phoneme /k/ will pass more activation to a competing phoneme /t/ via their shared manner and voicing features, than to a competing phoneme /d/, which only shares a manner feature. As such, a similar phoneme is more likely to be selected in place of the target phoneme than a dissimilar phoneme.

Two key theoretical claims are being made in this model. Firstly, it is quite explicitly assumed that subphonemic errors do not occur, even in aphasic speakers (Dell, 1986; Dell, Schwartz, et al., 1997). Secondly, the assumption that only the identity of the selected phoneme is passed on to subphonemic processes implies that no information cascades from phoneme level processes to subphonemic processes, and no information feeds back from subphonemic to phoneme level processes.

This section first summarises the key evidence used to motivate the claim that all speech errors originate at the phoneme level or above. Some problems with this evidence are then noted, based on results from the perceptual literature, and recent instrumental investigations of speech production. Finally, some instrumental investigations are summarised which have specifically sought to determine whether activation cascades from phonological encoding to subphonemic processes, and whether activation feeds back from subphonemic processes to phonological encoding. Alternative interpretations of some of the reported data are proposed, motivating the modelling of these results which this thesis presents.

2.3.1 Arguments for and against subphonemic speech errors

Arguments against subphonemic errors

The argument that all speech errors originate at the phoneme level or above is based on two key observations from early speech error investigations. Firstly, it has been argued there is no evidence for subphonemic errors. Subphonemic errors would theoretically result in a combination of articulatory features which either do or do not make a phoneme from the phoneme inventory of the speaker's language. Some research suggests that subphonemic errors which result in productions from outside

the speaker's phoneme inventory do not occur (Crompton, 1981), even when the speaker is suffering form jargon aphasia (Fromkin, 1971). Other research claims that subphonemic errors which result in phonemes from the speaker's inventory also do not occur. This position is based on the failure of some researchers to find many examples of errors which are unambiguous feature exchanges and cannot be accounted for as the exchange of phonemes in the context (MacKay, 1970; Nooteboom, 1969; Shattuck-Hufnagel, 1979; Shattuck-Hufnagel & Klatt, 1979). Fromkin (1973) provides an example of an unambiguous subphonemic error, shown in example 3a, where the mistakenly produced phonemes [g] and [p] were not in the phonemic context. However, whilst example 3b could be described as the replacement of a bilabial place of articulation with a velar place of articulation, resulting in an intended /b/ being produced as a [g], it could also be characterised as the replacement of the intended /b/ with the onset of *"gave"*, [g].

(3a) *"clear blue sky"* → *"glear plue sky"* (Fromkin, 1973)
(3b) *"gave the boy"* → *"gave the goy"* (Fromkin, 1973)
(3c) *"steak and potatoes"* → *"spake and tomatoes"* (Fromkin, 1973)

Secondly, the phenomenon of phonetic accommodation, where misordered phonemes are realised in a manner appropriate to their new environment, further supports the suggestion that errors occur at the phoneme level or above and that the articulatory plan is constructed based on phoneme level output. For instance, in example 3c (Fromkin, 1973), an intended /t/ was moved from a non-word-initial position, where it would not be aspirated. However, in the word-initial position to which the phoneme was displaced, the [t^h] was produced with aspiration, such that the final word did not sound like *"domatoes"*.

Problems with claims that subphonemic errors do not occur

However, these claims stem from transcribed evidence. Transcribing speech errors requires the listener to process the speech using their perception system, and results from the perceptual literature throw serious doubt on the reliability of this data as a record of human word production. It is well known that the development of perceptual skills for a child's native tongue coincides with a large decrease in ability to identify phonetic differences which are irrelevant to their language (Werker & Tees, 1984). The adult perceptual system has a strong tendency to classify input into phonemic categories (e.g., Buckingham & Yule, 1987), such that a transcriber would have difficulty in detecting productions outside his or her phonemic inventory.

Mispronunciations are also frequently ignored, especially those which involve a difference of only one feature (Cole, 1973; Marslen-Wilson & Welsh, 1978), rendering single feature errors much harder to detect. The skill of human perceptual processes in enabling efficient and functional comprehension is so extreme that listeners are even unable to detect when phonemes have been entirely replaced with a cough (Warren, 1970). It would therefore seem that the exquisite abilities of the human perceptual system in comprehending language make it an extremely inaccurate tool with which to gather the precise speech production data which is required to determine whether subphonemic errors occur and evaluate theories of information flow between phonological and phonetic processing stages.

Recent research has therefore turned to instrumental measures of the speech production system. Some instrumental investigations measure acoustic properties of the utterance, such as voicing onset time (VOT) or percent voicing (e.g., Frisch & Wright, 2002; Goldrick & Blumstein, 2006; McMillan, 2008). Others measure the movement of various articulators, including the tongue, lips and velum. Methods for capturing articulatory information include electropalatography (EPG), which measures tongue to palate contact (e.g., McMillan, 2008; McMillan et al., 2009); ultrasound, which captures an image of the shape of the tongue (e.g., McMillan, 2008; Pouplier, 2008); electromagnetic midsagittal articulometry (EMMA), which permits tracking of movement of points in the vocal tract through attachment of small transducer coils to the participant (e.g., Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007; Pouplier, 2007); electromyography (EMG) which measures muscle activity by means of an electrode inserted into the tongue or lips (e.g., Mowrey & MacKay, 1990); and X-ray, which can be used to obtain images of the articulators (e.g., Boucher, 1994). All of these investigations show that some productions exhibit intrusions of partial voicing or devoicing, or unexpected muscle activation or articulator movement, in both cases associated with a competing phoneme. Errors are reported to vary from slight intrusions to intrusions of a level associated with the production of a phoneme other than the phoneme intended (Frisch & Wright, 2002; Goldstein et al., 2007; Mowrey & MacKay, 1990; Pouplier, 2007). As such, these instrumental results strongly argue against a view in which subphonemic errors do not occur.

2.3.2 Instrumental investigations of information flow between phonological and subphonemic processing stages

Instrumental studies clearly offer an opportunity for us to gather huge amounts of fine-grained evidence from word production processes. In this thesis, we focus on two recent studies which have sought to capitalise on this methodology to address the question of information flow between phonological encoding and subphonemic processing stages. Specifically, Goldrick and Blumstein (2006) present acoustic evidence which they claim provides evidence for cascading from phonological encoding to subphonemic processes, and McMillan (2008) reports data in support of feedback from subphonemic processes to phonological encoding.

This section describes the evidence found in these studies, and conclusions drawn from these results. Some alternative explanations for some of the evidence are proposed. Finally, a plan for the work in the rest of this thesis is outlined, in which four versions of Dell's (1986) model spanning phonological encoding and subphonemic processes are implemented, and their abilities to account for this evidence investigated.

Goldrick and Blumstein's (2006) evidence for cascading from phonological encoding to subphonemic processes

Goldrick and Blumstein (2006) presented acoustic measurements of onset consonant productions. Goldrick and Blumstein (2006) asked participants to produce tongue twisters such as *"keff geff geff keff"*, where words in the tongue twister differed only by the voicing of the onset. The results demonstrated that when participants attempted to produce /k/s which were identified as sounding like [g]s, these [g]s were less voiced than intended /g/ onsets identified as [g]s. In other words, there was a trace of the intended voiceless phoneme /k/ on an errorful production of the voiced phoneme [g]. Traces were also observed for productions of voiced phonemes.

Goldrick and Blumstein (2006) argued that their findings could not be accounted for by noise at a subphonemic level, as there would be no reason for subphonemic noise to systematically affect only phonemes which were selected in error. Therefore, they claimed, a model with *cascading from all phonemes* to subphonemic levels is required. According to Goldrick and Blumstein's (2006) account, activation from the intended yet unselected phoneme cascades to articulation, even when noise in the model causes another phoneme to be selected. When a /k/ in a tongue twister is pronounced as [g], the phonemic representation of /k/ is relatively active (because

it was the originally intended onset) and this activation cascades, affecting the articulation of the errorful [g]. If a /g/ is intended and selected, there will be relatively little activation to cascade from the /k/. In a model without cascading from the unselected onset, the target /k/ cannot affect articulation, and the resulting output for an errorful [g] would not differ from an intended [g].

In support of this explanation, Goldrick and Blumstein (2006) reported a post-hoc analysis of the influence of lexicality on traces. In an account where there is feedback from phonological encoding to lexical selection (as argued for in section 2.2.2), the phonological form of the target will activate the lexical representation of the word error outcome, and this activation will in turn be conveyed to the phonological form of the word outcome. For example, in the error *"kess"* → *"guess"*, /gɛs/ will receive activation from *guess*. In contrast, nonword outcome errors, such as /gɛf/ in the error *"keff"* → *"geff"*, will receive no such activation. Goldrick and Blumstein (2006) assume that a mechanism exists by which a more activated /g/ will have a suppressive effect on the output from the /k/ representation. In this way, the intended /k/ will have less effect on the final production, such that an unintentionally produced [g] in the word outcome error *"kess"* → *"guess"* would have a smaller VOT trace of the intended /k/, than an unintentionally produced [g] in the nonword outcome error *"keff"* → *"geff"*. Whilst Goldrick and Blumstein's (2006) voiceless error outcome materials were not suitable for a lexicality analysis, as nearly all expected errors were non-lexical, their analysis confirmed that for voiced error outcomes, lexical errors demonstrated smaller traces than non-lexical errors as predicted by the cascading account.

Alternative less interactive accounts of Goldrick and Blumstein's (2006) evidence

However, there are two possible ways in which less interactive models would be able to account for Goldrick and Blumstein's (2006) findings. Each of these arguments is outlined in detail below.

The first alternative explanation is based on the premise that noise at a subphonemic level may indeed still be able to explain the basic trace results. By this account, *no cascading from phonemes* would be required. As observed by Goldrick and Blumstein (2006), there is no reason to believe that noise at the subphonemic level would affect incorrectly selected phonemes more than correctly selected phonemes. However, what is transcribed as an incorrectly selected phoneme may in fact be a correctly selected phoneme which has been affected by noise at a subphonemic level. For example, a speaker may intend to produce a /k/ and correctly select the

/k/ phoneme. Subphonemic noise may then lead to the [k] phoneme being realised as a [g]. Because the subphonemic level would retain activation from the [k], the errorful [g] would be more voiceless than a correctly produced [g].

Goldrick and Blumstein (2006) argued that if errors in tongue twisters were caused by noise at a subphonemic level, it would follow that the voicing of tokens identified as correct in a tongue twister task would be more variable than the voicing of tokens produced in a control task. In a post-hoc analysis presented by Goldrick and Blumstein (2006), no such difference is found. This is a null result, but Goldrick and Blumstein (2006) use it to argue against any account based on subphonemic noise. However, an account which explains errors in tongue twisters as being due to extra noise at the phoneme level, and assumes that activation cascades to the subphonemic level, should also predict that tongue twisters will lead to more noise at the subphonemic level. Goldrick and Blumstein's (2006) post-hoc null result therefore does not clearly differentiate between these different accounts.

Importantly, if errors were due to noise at the subphonemic level, there would be no way for activation at the lexical level to influence subphonemic processes. On this basis, we expect a model which captures Goldrick and Blumstein's post-hoc demonstration of a lexical effect on VOT to require cascading from selected phonemes.

The second alternative explanation assumes that the activation level of the selected phoneme may be lower if it has been selected in error. Such an account would require *cascading from selected phonemes only*. More specifically, an intended phoneme will receive activation from higher-level processes, and will therefore on average be more activated when selected than a phoneme which is activated through noise. An erroneously selected and therefore less activated /g/ may activate its subphonemic features less strongly than a correctly selected and therefore more activated /g/, such that an erroneously selected /g/ may result in a less voiced production. By definition, a less voiced production is more voiceless, such that a trace of the intended /k/ would be present in the final articulation without any activation having been transmitted from the /k/ at the phonemic level.

This account also permits explanation of the lexicality effect on traces found by Goldrick and Blumstein (2006). We follow Goldrick and Blumstein's (2006) assumption that an erroneously selected /g/ in the lexical outcome condition, for example on the error *"kess"* → *"guess"*, will be more activated than an erroneously selected /g/ in the non-lexical outcome condition, for example in the error *"keff"* → *"geff"*,

due to extra activation conferred upon the /g/ phoneme in the lexical outcome condition by the associated representation *guess* at the lexical level. In a model which permits cascading from selected phonemes, the extra activation from the erroneously selected lexical outcome /g/ will be transmitted to the subphonemic level, rendering the production more voiced. This will therefore result in a smaller trace of the voiceless /k/ in the lexical condition than in the non-lexical outcome condition.

Quantifying articulatory results of high level manipulations using the delta method

However, McMillan (2008) provides data which suggests that an even more interactive model than the model proposed by Goldrick and Blumstein (2006) is required. In these studies, McMillan (2008) uses the *delta method* introduced by McMillan et al. (2009). Whilst McMillan et al.'s (2009) results themselves do not place great constraints on phonological encoding to subphonemic information flow, they do demonstrate the effect of lexicality on low level articulatory measurements. Here, we outline the delta method, and at the same time provide a brief summary of McMillan et al.'s (2009) findings.

McMillan et al. (2009) measured tongue-to-palate contact using EPG in a Word Order Competition (WOC) task (Baars & Motley, 1976), which is designed to elicit onset errors. In the WOC task, participants are rapidly presented with pairs of nonwords, such as *gope doof*. These pairs are then replaced by arrows, pointing either left or right. If the arrow points right, participants should speak the words in the same order as seen on the screen, producing *"gope doof"*. If the arrow points left, participants should reverse the order, producing *"doof gope"*. Most fillers are followed by rightwards pointing arrows, cueing production in the correct order, whereas targets are always followed by leftwards pointing arrows, cueing reversal, and potentially causing an onset error, such as *"goof dope"*.

For half of the target items in McMillan et al.'s (2009) study, an onset error resulted in a lexical outcome, such as *"doof gope"* → *"goof dope"*, whereas for the other half, onset errors were non-lexical, as in *"doove gobe"* → *"goove dobe"*. All onsets in the target items were stops. Onsets in a pair always differed in place of articulation, where one onset was alveolar, and the other was velar. In half the pairs, onsets also differed in voicing.

Articulations in the lexical and non-lexical condition were compared using McMillan et al.'s (2009) delta method. The delta method permits a similarity value to

be calculated for two measurements of articulation. The method is based on a calculation of Euclidean distance. For simple one dimensional measurements such as VOT, this reduces to a calculation of the absolute difference. Measurements from EPG or ultrasound are more complex. In both cases, measurements of a single onset articulation can be characterised as a collection of vectors, where a vector represents tongue contact at each of the electrodes of an EPG palate, or the pixel greyscale values of an ultrasound image, and the size of the collection depends on how long the articulation took (or in other words, how many palate contact vectors or ultrasound images were collected). The delta method provides a transformation which standardises the length of two articulations, so that two collections of vectors are the same size, and calculates an average Euclidean distance between the vectors in the two collections. These two steps render the delta method sensitive to both differences in timing, and differences in location of palate contact (in EPG) or tongue shape (in ultrasound). Further detail on the calculation is provided by McMillan et al. (2009).

To compare articulations in the lexical and non-lexical conditions, reference EPG measurements were first obtained for each place of articulation. EPG recordings were taken from productions of the target phrases followed by a rightwards pointing arrow, cueing participants to produce the phrase in the presented order, so that participants were unlikely to make an error. The reference measurement was determined by calculating an average EPG recording for each of the two places of articulation.

For each of the EPG recordings of the experimental items, the delta method was used to calculate both the distance to the reference measurement for the target place of articulation, and the distance to the reference measurement for the competing place of articulation. For an intended production *"doof gope"*, the target place of articulation for the first onset /d/ would be alveolar, whereas the competing place of articulation for the first onset would be velar. It was found that in the lexical condition, articulations of onset phonemes are significantly more like reference measurements for the competing place of articulation than they are in the non-lexical error outcome condition. However, no significant difference between the two conditions was found for similarity of articulations to the reference measurement for the target place of articulation.

These results are very important as they demonstrate that it is possible to investigate the effect of high level variables on word production using instrumental measurements of articulation. Even more crucially, they introduce a quantitative

method free of transcriber bias for analysing such articulations. Notably, productions in McMillan et al.'s (2009) study were not categorised as either erroneous or error free. With very few exceptions for problematic recordings (e.g., recordings where no full closure between tongue and palate was recorded), all productions in both conditions were analysed, helping address the frequent problem of paucity of data in experimental speech error investigations. Most importantly, data was obtained entirely from instrumental measurements of the word production system, with no influence of the experimenter's language comprehension system, making this a pure record of human speech production.

McMillan et al. (2009) argued that this result was evidence for feedback from phonemes to words. They further made the reasonable suggestion that these results could be explained in a model in which activation cascades from unselected phonemes, as proposed by Goldrick and Blumstein (2006). In a model with feedback from phonemes to words and cascading from all phonemes, the extra activation conveyed to the competing phoneme /g/ in the lexical outcome condition would cascade to the feature layer, regardless of whether /g/ was selected or not.

However, we note that these results can be explained in any model of information flow from phonological encoding to subphonemic processes. The lexical bias result would suggest that the competing onset phoneme is selected more often in the lexical condition than in the non-lexical condition. Even in a model with no cascading from phonological encoding, more frequent production of the competing onset in the lexical condition would lead to an average articulation closer to the reference measurement for the competing place of articulation than the average articulation in the non-lexical condition. It is not clear that any of the models under consideration would predict that the distance of articulations from the target should not be affected by lexicality, regardless of interactivity. However, it is also not clear what the power of this experiment is, and so explanation of this null result is not prioritised in the current thesis. Finally, the observation that intermediate articulations occur also does not distinguish between models, as such productions could be explained either by activation cascading to the subphonemic level, or noise at the subphonemic level.

McMillan's (2008) evidence for feedback from subphonemic processes to phonological encoding

The experiment described by McMillan (2008) applying McMillan et al.'s (2009) methodology does impose strong constraints on information flow from phonological

encoding to subphonemic processes, however. McMillan (2008) reports articulatory and acoustic measurements of onset consonant productions from tongue twister productions. Acoustic information was obtained by measuring VOT. Articulatory information was obtained in one study using EPG and in another study, by using ultrasound. An advantage of ultrasound over EPG is that partial tongue movements which do not reach the palate (as found by e.g., Goldstein et al., 2007) can be taken into account (McMillan, 2008).

In the analyses focused on here[1], McMillan (2008) again used materials with alveolar and velar stop onsets. Place and voicing features of the onsets were varied orthogonally. This resulted in four conditions, in which onset pairs were either the same (e.g., *"teff teff"*), or differed in place feature (e.g., *"teff keff"*), or differed in voicing feature (e.g., *"teff deff"*), or in the final condition, differed in both place and voicing feature (e.g., *"teff geff"*). Using the tongue twisters in which the onset consonants were the same (e.g.. *"teff teff"*), reference acoustic and articulatory measurements were obtained for each onset phoneme. The distance from these reference measurements to the acoustic and articulatory measurements in the other conditions was then calculated, again using the delta method introduced by McMillan et al. (2009).

The articulatory results showed that articulatory measurements were further from the reference (e.g., *"teff teff"*) when the onsets differed in place (e.g., *"teff keff"*) than when the onsets differed in both place and voicing (e.g., *"teff geff"*). Similarly, the acoustic results showed that acoustic measurements were further from the reference when the onsets differed in voicing (e.g., *"teff deff"*) than when the onsets differed in both place and voicing (although this result failed to reach significance in the ultrasound study).

McMillan's (2008) findings suggest that there is *feedback from subphonemic representations* to phonological encoding. With subphonemic level to phonemic level feedback, competing phonemic representations which share more features (and therefore differ by fewer features) will receive more activation via feedback from

[1] A further analysis of the ultrasound data also included materials with alveolar fricative onsets. Voicing data was not considered in this analysis, as voicing is measured differently for fricatives and stop consonants. Articulatory results from this analysis were surprising however, showing that articulations were most dissimilar when onsets differed in place, manner and voicing. Note though that as the English phonemic inventory does not include velar fricatives, these materials were not counterbalanced. McMillan (2008) observes that it is not clear that the results from this analysis are not due to this imbalance in the materials, and as suggested by McMillan (2008), we postpone consideration of this result until it has been replicated in a language in which all three features can be crossed in a balanced design.

subphonemic representations. For example, when the target phoneme is /t/, a competing phoneme /k/ will receive more activation than a competing phoneme /g/, due to the voicing feature shared by /t/ and /k/. Activated phonemic representations will then pass activation to their component subphonemic representations, including the competing voicing or place representation. For example, for a target production /t/, the target place feature is *[alveolar]*. However, both /k/ and /g/ would activate the competing place feature *[velar]*. Because similar phonemes receive more activation from the target phoneme, the competing voicing or place subphonemic representation will receive more activation when a more similar phoneme is competing. Following the example, /k/ would pass more activation to the competing place feature *[velar]* than /g/ would. Such processing mechanics would explain McMillan's (2008) results, as measurements of articulation will be further from the reference articulation when the competing onset differs in place but shares a voicing feature than when the competing onset differs in both place and voicing; and equally, measurements of voicing will be further from the reference voicing measure when the competing onset differs in voicing but shares a place feature than when the competing onset differs in both place and voicing.

The transcribed phonological similarity effect and feedback from subphonemic representations to phonological encoding

In Dell's (1986) original model, feedback from the layer of features to the phoneme level is assumed in order to explain the transcribed phonological similarity effect. However, output at this model is at a phoneme level. In section 2.3.1, we argued that the transcribed evidence used to motivate this design decision is unreliable in the light of results from the perceptual literature and instrumental investigations. In fact, instrumental investigations clearly demonstrate that subphonemic errors do occur. In a model in which the possibility of subphonemic errors is no longer ruled out, and output is determined at a subphonemic level, subphonemic to phonemic feedback is no longer required to explain the transcribed phonological similarity effect. Rather, the effect can be explained simply because fewer subphonemic representations must be misactivated for a production to sound like a similar phoneme, than for a production to sound like a dissimilar phoneme. For example, for a target phoneme /t/ to be produced as a [k], only the place feature must be misactivated. However, for a /t/ to be produced as a [g], both the place and voicing features must be misactivated.

Table 2.2: Activation flow characteristics of the four proposed models of information flow between phonological encoding and subphonemic processes

Model	**Information from phonological encoding**			**Feedback from subphonemic representations**
	Identity of selected phoneme	*Activation from selected phoneme*	*Activation from unselected phonemes*	
No casc	✓			
Casc from sel	✓	✓		
Casc from all	✓	✓	✓	
Feedback	✓	✓	✓	✓

However, in a model without feedback from subphonemic processes to phonological encoding, where the phonological similarity effect was explained solely by subphonemic noise, there would be no reason for articulatory measurements to be closer to the reference measure when onsets differ in both place and voicing in comparison to when onsets differ just in place of articulation; rather, the amount of noise on the place feature would remain the same. Similarly, for the acoustic measurements, there is no reason that noise on the voicing features should decrease when onsets differ in both place and voicing in comparison to when onsets differ just in voicing.

This makes McMillan's (2008) results particularly interesting, as with this in mind, they appear to be the only evidence for feedback from subphonemic processes to phonological encoding in the literature, again emphasising the usefulness of instrumental measurements of word production.

Modelling information flow between phonological encoding and subphonemic processes

The core goal of the current thesis is to build upon this recent evidence with computational simulations. Specifically, it aims to extend Dell's (1986) influential model past the phoneme level, and compare the ability of different models of information flow between phonological encoding and subphonemic processes to account for this new data.

In these simulations, we examine the behaviour of four models of information flow between phonological encoding and subphonemic processes which were referred to in this section: *no cascading from phonemes*, *cascading from selected phonemes only*, *cascading from all phonemes*, and finally *feedback from subphonemic representations*. Table 2.2 depicts the differences in activation flow between these models.

These simulations have three main goals. Firstly, we aim to simulate acoustic and articulatory measurements of word production within the framework of Dell's (1986)

model for the first time. All previous simulations using Dell's (1986) model have only modelled transcribed evidence. Whilst the current simulations will include transcribed evidence, we also simulate VOT results (Goldrick & Blumstein, 2006; McMillan, 2008) and EPG and ultrasound results (McMillan, 2008; McMillan et al., 2009).

Secondly, we aim to demonstrate that contrary to Goldrick and Blumstein's (2006) claims, cascading from all phonemes is not required to account for their data. We predict that a model with no cascading from phonemes can account for the basic VOT trace data, and that a model with cascading from selected phonemes only can account for both the basic VOT trace data and the lexicality effect on VOT traces from Goldrick and Blumstein's (2006) post-hoc analyses.

Thirdly, we aim to show that while any model of information flow between phonological encoding and subphonemic processes can account for the transcribed phonological similarity effect when output at the subphonemic level is assumed, the results reported by McMillan (2008) can only be accounted for in a model including feedback from subphonemic representations to phonological encoding.

Alongside these simulations, we hope to show that McMillan et al.'s (2009) results can also be modelled in this framework, although they do not impose strong constraints on models of information flow between phonological encoding and subphonemic processes.

Our full set of predictions for these models are shown in table 2.3. In these predictions, we assume feedback from phonological encoding to lexical selection, as argued for in section 2.2.2; without this assumption, we do not expect that any of the models would be able to account for lexical bias effects.

2.3.3 Summary

In this section, the evidence used to motivate the assumption of phoneme output in Dell's (1986) original model was examined. This evidence all stems from speech error studies in which the experimenter has transcribed what they heard. It was argued that in the light of results from the perceptual literature, transcribed evidence cannot be relied on to make this sort of judgement, and that the conclusions drawn from the transcribed evidence are not convincing. Furthermore, results from studies using instrumental measurements of speech production are not in line with a view which assumes that subphonemic errors do not occur.

Table 2.3: Predictions of the ability of different models of information flow between phonological encoding and subphonemic processes to account for empirical data (assuming feedback from phonological encoding to lexical selection).

Model	*transcribed LB*	*transcribed PS*	*G&B 2006 traces*	*G&B 2006 trace LB*	*MMea 2009 delta LB*	*MM 2008 delta PS*
No casc	✓	✓	✓	×	✓	×
Casc from sel	✓	✓	✓	✓	✓	×
Casc from all	✓	✓	✓	✓	✓	×
Feedback	✓	✓	✓	✓	✓	✓

Key:
LB = lexical bias, PS = phonological similarity, G&B 2006 = Goldrick and Blumstein (2006), MM 2008 = McMillan (2008), MMea 2009 = McMillan et al. (2009)

✓ = predicted to be able to account for evidence
× = predicted not to be able to account for evidence

Grey boxes indicate that our prediction does not match the standard claim in the literature.

The section then considered evidence from two instrumental studies which have examined the question of information flow between phonological encoding and subphonemic processes. Firstly, Goldrick and Blumstein (2006) presented evidence of VOT traces on erroneously produced phonemes, from which they claimed that *cascading from all phonemes* to subphonemic processes is required. Two alternative accounts of this data were presented. It was argued that the first account, in which there is *no cascading from phonemes* to subphonemic processes, can account for the presence of traces. The second account, in which there is *cascading from selected phonemes only*, can account for both the presence of traces, and a post-hoc result demonstrating an effect of error outcome lexicality on traces.

An innovative method introduced by McMillan et al. (2009) for quantifying articulatory measurements so that they can be compared between different experimental conditions using standard statistical approaches was then outlined. McMillan et al.'s (2009) results demonstrating the effects of error outcome lexicality on articulation were summarised, although it was noted that these results do not constrain models of information flow from phonological encoding to subphonemic processes. McMillan's (2008) results on the other hand strongly suggest that *feedback from subphonemic representations* is required. It was highlighted that these results are particularly interesting as the classic transcribed phonological similarity effect is no longer evidence for feedback from subphonemic processes to phonological encoding, once the assumption of output at the phoneme level is removed.

Finally, this section outlined a plan to run simulations to investigate the four proposed models of information flow between phonological encoding and subphonemic processes, and verify the theoretical claims made about their ability to account for

the various sets of data. In simulating this data, the modelling work outlined would also comprise the first example of acoustic and articulatory data being modelled within the framework of Dell's (1986) model of word production.

2.4 Investigating different information flow options in a spreading activation model: the parameter problem

Spreading activation models such as the one suggested by Dell (1986) require a number of parameters to be set in order for the activation of representations in the models to be calculated. Representations are connected to other representations in stages directly preceding and following the current stage. Representations receive activation from connected representations at the higher level stage, depending on how much cascading from the higher level stage is assumed, and how many stages of word production have been completed. Representations may also receive activation from connected representations at the lower level stage, if feedback from that stage is assumed. The amount of activation received from each connected representation at the previous stage is equivalent to the level of activation of the representation at the previous stage, multiplied by the *forward connection strength*. Equally, the amount of activation received from each connected representation at the following stage is equivalent to the level of activation of the representation at the following stage, multiplied by the *feedback connection strength*.

Time in Dell's (1986) model is quantized into *steps*. At each timestep, the activation of a representation is calculated based on the sum of activation received from connected nodes, and the activation of the node at the previous timestep. Activation of a representation is subject to *decay*, applied at each timestep, such that only a proportion of the activation of the representation remains. Finally, noise also affects the level of activation of a representation, increasing or decreasing it by a random value which follows a normal distribution. There are two types of noise. The variance of the first, *activation dependent noise*, is greater when the activation level of a representation is higher. The variance of the second, *intrinsic noise*, is not affected by the activation level of a representation, and corresponds to a set level of background noise. Most simulations based on Dell's (1986) model at least include activation-based noise, and some utilise intrinsic noise as well.

After a pre-defined number of steps, an amount of activation known as a *jolt* is added to representations which have been selected for a slot in the frame. Representations

which are buffered for later production have an amount of activation known as a *prime* added to their activation level.

These eight parameters (forward connection strength, feedback connection strength, jolt, prime, decay, steps, activation-based noise and intrinsic noise) do not map on to human attributes in a way that prescribes their settings. Dell (1986) has suggested that a lower number of steps, and therefore shorter amounts of time between selection stages, could reflect faster speech rates. Other researchers have suggested links between certain parameter settings and aphasic damage to the word production system. For example, Rapp and Goldrick (2000) increase activation dependent noise at different representation levels to simulate damage at these levels. Dell and colleagues (Dell, Schwartz, et al., 1997; Foygel & Dell, 2000) have proposed that decreased forward and feedback connection strength or increased decay of activation may be responsible for aphasic error patterns. However, even if these hypotheses are accepted, it is not clear what the transformation is from speech rate or measured aphasic impairment to parameter setting. The only way to estimate the parameter setting is to choose a parameter setting, run the model, and determine the appropriateness of the selected parameter setting through comparison of the model's behaviour to relevant empirical results.

Furthermore, previous studies do not dictate a specific set of parameter settings. Table 2.4 shows all the parameter settings for simulations using models based on Dell's (1986) theory that we know of. As is clear from the table, the settings chosen vary substantially between studies. For example, Hartsuiker (2002) presents simulations closely based on Dell's (1986) original model. The parameters selected, based on a criterion that the model should be correct in 93% of segmental positions when words are generated in isolation, turn out to be rather different to the parameters used by Dell (1986). In particular, forward and feedback connectivity is much weaker, with forward connection strength set to 0.1 and feedback connections strength set to 0.05, in comparison to 0.3 for forward connection strength and 0.3 or 0.15 for feedback connection strength in Dell's (1986) model.

Similarly, the model presented by Dell, Schwartz, et al. (1997) is based on Dell's (1986) model of phonological encoding, with a few key differences, such as its focus on one word utterances, inclusion of a semantic level, absence of a featural level, and its intentionally small vocabulary, designed to aid variation of the connectivity and decay parameters to fit aphasic behaviour. The basic parameters used prior to parameter variation, selected to allow the model to exhibit the proportions of correct, semantic, formal, nonword, mixed and unrelated errors found in non-aphasic

Table 2.4: Parameter settings used in previous simulations based on Dell's (1986) theory. Numbers given are the key parameter settings used in the papers to simulate normal behaviour - variations on these settings are discussed in the text

	fwdConn	fbkConn	jolt	prime	decay	steps	actiNoiseSD	intrinNoiseSD
Simulations which made errors due to noise								
Dell (1986) (error types)	0.3	0.15	100	50	0.6	3, 4, 8	0.2	0
Martin, Dell, Saffran, and Schwartz (1994)	0.1	0.1	100	*N/A*[c]	0.4	8	0.18	0.01
Dell, Schwartz, et al. (1997)[a]	0.1	0.1	100[b]	*N/A*[c]	0.5	8	0.16	0.01
Foygel and Dell (2000)[d]	0.1	0.1	100[b]	*N/A*[c]	0.6	8	0.16	0.01
Rapp and Goldrick (2000)[e]	0.04	0.04	4[f]	*N/A*[c]	0.5[g]	8	0.1[h]	0
Goldrick (2006)	0.05	0.05	4[f]	*N/A*[c]	0.5[g]	8	0.1[h]	0
Hartsuiker (2002)	0.1	0.05	100	50	0.5	5	0.05	0
Oppenheim and Dell (2008)	0.2	0.2	1[h]	0.01[h]	0.4	4	0.68	0
Simulations which made errors by means other than noise (e.g., insufficient time for encoding, anticipatory or perseveratory bias)								
Dell (1986) (speech rate)	0.3	0.15	100	50	0.6	2, 3, 4, 5	0	0
Dell (1986) (SLIP simulation)[j]	0.3	0.3	60	30	0.4	2, 3, 4	0	0
Dell (1988) (SLIP simulation)[j]	0.18	0.18	60	30	0.4	2, 3, 4	0	0
Simulations which would theoretically make errors due to a stochastic selection rule								
Dell (1988, 1990) (frequency)	0.1	0.1	10	*N/A*[c]	0.2	1 to 22[k]	0	0
Dell and O'Seaghdha (1991, 1992) (mixed errors)	0.1	0.1	10	*N/A*[c]	0.4	1 to 10[k]	0	0
Dell and O'Seaghdha (1991, 1992) (time course)	0.1	0.1	1	*N/A*[c]	0.4	8	0	0

Key fwdConn = forward connection strength, fbkConn = feedback connection strength,

actiNoiseSD = factor by which the activation of a representation is multiplied to determine the standard deviation of activation-based noise

intrinNoiseSD = standard deviation of intrinsic noise

[a]Other papers using this model and its parameter settings: Dell and Gordon (2003); Dell, Lawler, Harris, and Gordon (2004); Ruml and Caramazza (2000); Ruml, Caramazza, Shelton, and Chialant (2000); Ruml, Caramazza, Capasso, and Miceli (2005); Schwartz, Dell, Martin, Gahl, and Sobel (2006)

[b]This is the jolt setting for the lexical level, at the start of phonological encoding. The jolt at the semantic level was 10.

[c]Models which do not produce multiple words in sequence do not have a prime parameter.

[d]Other papers using this model and specifically its decay setting: Dell et al. (2004); Schwartz et al. (2006)

[e]Other papers using this model and its parameter settings: Goldrick and Rapp (2002); Ruml et al. (2000, 2005)

[f]This is the jolt setting for the lexical level, at the start of phonological encoding. The jolt at the semantic level was 10.

[g]This is the decay level for most of the network, although decay at the concepts level was set to 0.7

[h]This is the lowest noise level specified, although noise is manipulated to higher values to simulate aphasic damage

[h]In this small model, the jolt is applied at the word level but the prime is applied at the phoneme level, to simulate the SLIP task

[j]Dell (1986) states that other combinations of connection strength, decay and timesteps are tested, but only results for the parameters shown are provided.

[k]Selection is not simulated in this model, and instead activation levels are plotted across time.

production, are really quite different to those used in Dell (1986) however. Again, the forward and feedback connection strength is much lower than in Dell's (1986) original investigation (0.1 for both forward and feedback connection strength in the study reported by Dell, Schwartz, et al., 1997), and Dell, Schwartz, et al. (1997) introduce the concept of intrinsic noise, whereas this is absent in Dell's (1986) original work. Dell, Schwartz, et al. (1997) also allow 8 steps per selection stage, whereas Hartsuiker (2002) allows 5.

Developments of Dell, Schwartz, et al.'s (1997) work then change the parameters further. Foygel and Dell (2000) state that a decay value of 0.6 is used in their model, instead of the 0.5 used by Dell, Schwartz, et al. (1997), but no reason is given for this change. Equally, Rapp and Goldrick (2000) state that they are building on Dell, Schwartz, et al.'s (1997) model, but their parameters are quite different, including much weaker forward and feedback connection strengths (0.04 instead of 0.1), a much smaller jolt size at the start of phonological encoding (4 instead of 100) and no intrinsic noise.

Further differences do exist between these models, such as the number of levels per selection stage, the number of words produced by a model, the use of resting activation levels, and indeed the evidence which a model is set up to simulate. However, no clear relationship between these differences and the difference in parameter settings is given by any of the modellers. It is also not clear whether these variations in parameter settings change the model's behaviour sufficiently to invalidate any assumptions a reader may wish to draw that these different simulations capture different aspects of a larger theory.

Given these problems, it is not clear what parameter settings should be chosen for the current simulation study. Would it be valid to compare different information flow options with one largely arbitrarily chosen set of parameter settings? This question reveals an underlying uncertainty - what are the effects of changing the various parameter settings on model behaviour?

The next two sections summarise what previous studies have shown about the effects of changing parameter settings, and how previous researchers have approached the problem of investigating different information flow options in Dell's (1986) spreading activation model.

2.4.1 The effects of manipulating parameters in the spreading activation model

Previous investigations give us some clues about the potential effect of manipulating parameters within models based on Dell's (1986) theory.

Connection strength

Dell, Schwartz, et al. (1997) investigated the effect of weakening forward and feedback connection strength in a model with semantic, lexical and phonological representations, with the aim of simulating aphasic error patterns (see also early work in this direction by Martin et al., 1994). They found that as connection strength was reduced from their chosen base value of 0.1, more nonword errors were produced, as well as unrelated word errors, where the produced word had no semantic or phonological relation to the target. Dell, Schwartz, et al. (1997) explain this result by suggesting that reducing connection weight results in activation patterns on different levels being inconsistent with each other. Nonword errors result from inconsistencies between the lexical level and the phonological level. Unrelated word errors result from inconsistencies between the semantic level and the lexical level (although arguably the reduced influence of the phonological level also plays a role here, causing unrelated rather than formal errors to be produced).

Foygel and Dell (2000) (see also Ruml et al., 2000, 2005) investigate the effect of independently manipulating the connection weight between the semantic and lexical level, and the connection weight between the lexical and phonological level. Again beginning with a base value of 0.1, decreased forward and feedback connection weights between the semantic and lexical level led to a high incidence of semantic, formal, mixed and unrelated errors, but not to an increase in nonword errors. Foygel and Dell (2000) explain that the reduced strength of the semantic to lexical connection means that lexical selection is largely random, but that the selected word is then encoded normally, such that very few nonwords are produced. Foygel and Dell (2000) also note that formal errors are promoted over entirely unrelated errors, as when the top-down semantic influence on lexical selection is reduced, the effects of the bottom-up phonological influence are relatively stronger, leading to more frequent misselection of formal competitors. In contrast, when the connection strength between lexical and phonological representations is decreased, phonological encoding cannot proceed correctly and nonword errors prevail.

Finally, Dell and Gordon (2003) investigated the results of reducing network wide connection strength on neighbourhood density effects, again in a model with semantic lexical and phonological representations. As noted in section 2.2.2, Dell and Gordon (2003) showed that in the spreading activation model, words are more accurately encoded in dense phonological neighbourhoods, as feedback from phonology activates neighbours which then provide extra activation support to the shared phonology. Extra activation of the shared phonology also provides further support to the original target word at the lexical level. Formal errors (i.e., the production of a neighbour instead of a target word) are predictably more common in denser neighbourhoods, and can result either from feedback activation of the neighbour at the lexical level, or misactivation of one phoneme at the phonological level. However, as semantic errors are by far the most common error overall due to activation of semantic neighbours at the lexical level, and so few formal errors are actually produced, this minor increase in formal errors does not overrule the overall accuracy boost.

In line with Dell, Schwartz, et al.'s (1997) results, Dell and Gordon (2003) found that reducing forward and feedback connection strength to 0.0033 rather than the original 0.1 caused a large increase in overall error rate. In addition, the dense neighbourhood accuracy boost was eliminated. The increase in errors can easily be explained as being due to weakened support for intended productions, from the semantic level to the lexical level, and from the lexical level to the phonological level. Equally, feedback support from the phonological level to the lexical level will be weakened, so that the effect of neighbourhood is reduced. Dell and Gordon's (2003) results appear to suggest that with weakened connection strength, accuracy is actually worst in the denser neighbourhoods, both for lexical selection and phonological encoding. This may be due to feedback interactions with neighbours further increasing the noise in the weak connection network. However, in Dell and Gordon's (2003) simulations, neighbourhood size is confounded with model size, as formal neighbours are simply added to the model to simulate denser neighbourhoods. The presence of extra nodes may therefore simply provide more opportunities for random errors in the denser neighbourhoods.

Several researchers also note that problems may be caused by choosing a connection strength which is too high given the chosen decay rate (Dell, 1988; Dell & O'Seaghdha, 1991, 1992; Schade & Berg, 1992; Shrager, Hogg, & Huberman, 1987; Waltz & Pollack, 1985). Shrager et al. (1987) use a generic spreading activation net to show that the connection strength must be less than the decay rate for activation

in the network to not rise without end because of reverberation in the network due to feedback. Dell (1988) further suggests that the connection strength should be less than half of the decay rate, although it is not clear where this figure comes from, and simulations of the SLIP task presented by Dell (1986) where the connection strength is 0.3 and the decay rate is 0.4 do not meet this constraint.

To recap, most studies investigating manipulations of connection weight suggest that reducing connection weight below a level believed to be appropriate for normal production leads to inconsistencies in activation patterns at different levels, a reduction in interactive effects, and an increase in errors overall. Another study abstractly examining the behaviour of spreading activation models suggests that increasing the connection weight past a certain point may also prevent the network from behaving reasonably.

Feedback connection strength

A few researchers have also investigated the effect of manipulating feedback connection strength only, although these studies involved either models with aphasic damage (Goldrick, 2006; Rapp & Goldrick, 2000; Ruml et al., 2000), or models with an unusual architecture (Hartsuiker, 2002).

In a model with a semantic, lexical and phonological level, Ruml et al. (2000) extended Dell, Schwartz, et al.'s (1997) investigations, examining the effect of further reducing feedback connection strength when connection strength was already reduced below the 0.1 used by Dell, Schwartz, et al. (1997). When feedback connection strength was reduced to a tenth of the forward connection strength, similar error patterns to those found by Dell, Schwartz, et al. (1997) were generated. However, with feedback strength reduced to 1/100 of the forward connection strength, the model produced only errors at all parameter settings. Similarly, when noise was introduced at the lexical and phonological levels instead of global connection weight being reduced, many errors were generated when feedback strength was low. Ruml et al. (2000) suggest that this result is due to the reduced effect of reinforcement, leading to pronounced effects of background noise.

Rapp and Goldrick (2000; Goldrick, 2006) explored the effect of manipulating feedback to a damaged level, and feedback from a damaged level, where damage was simulated by increased activation-based noise on representations (a standard deviation of 0.7 in Goldrick's 2006 simulations). Strong feedback to a damaged level led to a strong influence of lower levels on error patterns. Specifically, strong feedback

from the phoneme level to a damaged lexical level resulted in many formal errors at lexical selection. This was particularly true in Goldrick's (2006) simulations when feedback strength was higher than 0.05, which was the strength of the forward connections. Stronger feedback from the phoneme level to a damaged lexical level also increased non-word production, especially when words were less strongly selected due to a reduced jolt (Rapp & Goldrick, 2000). This is because feedback from the phoneme level to the lexical level activates neighbours of selected words, which then activate their phonology, sometimes resulting in a combination of phonemes being selected which do not form a word.

Strong feedback from a damaged level to a previous level also disrupted processing at the previous level. For example, feedback from a damaged lexical level to the semantic level resulted in large numbers of errors at semantic selection. In Goldrick's (2006) simulations, this was particularly true when feedback was set to 0.05 or higher, which was the strength of the forward connections.

Finally, Hartsuiker (2002) investigated the effect of manipulating feedback strength (from 0 to 0.1) in a model including syllable shapes (e.g., CV, or CVC). In Hartsuiker's (2002) model, words are connected to syllables, and syllables are connected to both syllable shapes and phonemes, which both feed activation back to the syllable representations. The shape of the produced syllable is determined by the selected syllable shape, and the content of the syllable is determined by the phonemes selected. No selection occurs at the syllable level, and syllable shapes and phonemes are selected at the same time. Hartsuiker (2002) found that stronger feedback from phonemes to syllables favour bigger syllables (e.g., CVC instead of CV), as syllables containing more phonemes receive more activation. Stronger feedback from syllable shapes to syllables favour syllables with more common syllable shapes (e.g., in Spanish, CV instead of CVC), because syllable shapes which are connected to more syllables receive more activation.

These results show that feedback manipulations can affect processing in a number of ways. Ruml et al.'s (2000) results underline the role that feedback can play in reinforcing intended productions, and the errors that can result from removing this reinforcement by reducing the strength of feedback. In contrast, the results presented by Rapp and Goldrick (2000; Goldrick, 2006) demonstrate that too strong an influence from lower level processes can be problematic, especially if either the lower level process or the process it is feeding back to is impaired. Lastly, Hartsuiker's (2002) investigations serve as a reminder that feedback particularly increases the activation of representations which are connected to many other representations

(Dell, 1986). The higher number of connections may be due to greater size of the representation (as in the case of the bigger syllables), but is often due to higher frequency of the representations (as in the case of the more frequent syllable shapes).

Jolt size

In models which produce one word only (in which no upcoming words are primed), some researchers have investigated manipulating jolt size. Rapp and Goldrick (2000; Goldrick, 2006) focus on the role that jolt size plays in a model with cascading from unselected representations. They argue that jolt size determines how strongly a representation is selected. In a model with a single unit concept level, semantic feature level, lexical level and phoneme level, Goldrick (2006) (see also Rapp & Goldrick, 2000) has shown that when the lexical level is damaged by setting the standard deviation of activation-based noise at that level to 0.7, decreasing the jolt given to selected lexical representations below the standard setting of 4 results in increasing disruption of the phonological encoding process, evidenced by increasing proportions of non-word responses. Lower jolt sizes mean that the activation cascading from unselected words is high, proportional to the activation cascading from the selected word. The phonology of unselected words therefore becomes proportionally very active, and random selections of phonemes begin to occur.

Rapp and Goldrick (2000) have also demonstrated that when the lexical level is damaged, comparatively high jolt sizes (in their simulations, a jolt size of 10) do not permit a strictly feedforward model with cascading from the lexical level to phonological encoding to exhibit a mixed error effect. With such a parameter setting, activation from unselected but activated semantic competitors at the lexical level can no longer transfer effectively to the phoneme level. Rapp and Goldrick (2000) suggest that with a high enough jolt size, a model with cascading from unselected representations could be made to generate the same output patterns as a model in which no activation is transmitted from unselected representations.

Dell, Schwartz, et al. (1997) make a theoretical comment on the role of the jolt parameter, although no simulations are provided to back this assertion up. Dell, Schwartz, et al. (1997) note that the jolt size "sets the activation scale" in the network (Dell, Schwartz, et al., 1997, pp. 813). They claim that there is therefore an inverse relationship between the effect of manipulating the size of the jolt and the level of intrinsic noise in the network, such that a 50 fold increase in jolt size is equivalent to a 50 fold decrease in intrinsic noise.

To summarise, jolt size can be seen as an index of selection strength (Goldrick, 2006; Rapp & Goldrick, 2000). However, jolt size also plays an important role in determining the amount of activation in the network overall, which affects the influence that other parameters such as intrinsic noise have on the behaviour of the network (Dell, Schwartz, et al., 1997).

Prime

Very few simulations have investigated the effect of manipulating the amount of activation given to buffered representations, known as the prime. Many simulations based on Dell's (1986) model have focused on production of one word only (e.g. Dell, Schwartz, et al., 1997; Foygel & Dell, 2000; Rapp & Goldrick, 2000; Goldrick, 2006, amongst many others) and therefore do not include a prime parameter. In the two published simulations which do produce more than one word, the prime is set to be half of the jolt (Dell, 1986; Hartsuiker, 2002). However Dell (1986) explicitly states in his SLIP simulation that this ratio has "no claim to any motivation other than [his] belief that it would be a good value" (Dell, 1986, pp. 306). The only investigation into manipulation of the prime parameter was carried out by Dell (1986) in his phonological encoding simulation. Dell (1986) reports that the chance of an exchange or anticipation being generated could be increased by using a larger amount of priming activation. Whilst this makes intuitive sense, as the extra activation provided to an upcoming representation will make it more likely that it is selected early, no simulation data is provided relating to this investigation.

Decay

Manipulations of decay have been examined in some detail, for single word production models at least. Most of this work is oriented around Dell, Schwartz, et al.'s (1997) proposal that increasing decay may provide an avenue for simulating aphasic damage (see also early work in this direction by Martin et al., 1994). In a model with a semantic, lexical and phonological level, Dell, Schwartz, et al. (1997) show that increasing decay above the default value of 0.5 increases generation of mixed, formal and semantic errors. Decay based errors show a clearer influence of interactivity between levels than errors caused by weakened connection strength, resulting in related although incorrect productions. Descriptions of simulation results provided by Ruml et al. (2000) suggest that formal errors can be promoted if decay is increased at both the semantic and the lexical level, but not the phonological level,

due to the increased influence of the phonological level via feedback to the damaged levels.

Taking Dell, Schwartz, et al.'s (1997) lead, Dell and Gordon (2003) investigated the effect of increasing decay throughout a network on the previously reported accuracy boost for words in denser neighbourhoods. Whilst overall error rate increased, the accuracy boost was not affected by this manipulation. As decay does not interfere with connectivity, the feedback mechanism which affords the accuracy boost to words in denser neighbourhoods was not disrupted.

However, Dell, Schwartz, et al. (1997) emphasise that the focus on the decay parameter in these investigations was for ease only, and suggest that other parameters affecting the amount of noise in the network could have been manipulated to the same effect. For example, they provide simulation results demonstrating that increasing intrinsic noise resulted in behaviour very similar to when decay was increased and connection strength slightly reduced.

Dell (1986) also investigated the effect of decay rate on productions of multiple words. He reports that at slower decay rates, more exchanges and perseverations are produced, at the expense of anticipations, more of which are transformed into exchanges. Decay rate manipulations in this investigation therefore appear to mostly affect the second word. Production of an exchange *"big fun"* → *"fig bun"* requires the /b/ which was intended for the onset of the first word but was not selected, to maintain the activation it was given for production of the first word such that it is sufficiently activated to be selected at the second word. As lower decay rates lead to better maintenance of activation, it follows that they will also increase exchange rates. Production of a perseveration *"big fun"* → *"big bun"* requires the /b/ which was intended for the onset of the first word and was selected and suppressed, to be reactivated by word nodes it is connected to. It is conceivable that a lower decay rate will help here too, as the connected word nodes will retain higher activation for longer, and will therefore have more activation to pass back to the suppressed /b/. Dell (1986) further notes, however, that too low a decay rate causes the model to generate chains of perseverations leading to "complete gibberish" Dell (1986, pp. 300), which would fit in with Shrager et al.'s (1987) report that low levels of decay combined with high connection strengths result in the network's behaviour being too strongly affected by events in the past. However, as with the investigations into prime manipulation reported by Dell (1986), no precise data is provided from these investigations.

To summarise, investigations in models which produce single words have shown that very high levels of decay lead to the models generating many errors, although these errors tend to be related to the target words, reflecting the unimpaired connectivity between layers of representation (Dell & Gordon, 2003; Dell, Schwartz, et al., 1997; Ruml et al., 2000). Investigations in a model which produces multiple words (Dell, 1986) focused on the results of reducing decay on second word errors, suggesting that lower decay levels lead to increased perseveration and exchange rates, at the expense of anticipations. Reducing decay too far resulted in nonsensical productions due to chains of perseverations. No simulations report the effect of low levels of decay on single or initial word productions (although Shrager et al.'s 1987 results showing that the decay rate must be greater than connection strength for the network to behave reasonably should be remembered). Similarly, no simulations report the effect of very high levels of decay on productions of non-initial words.

Steps

The parameter setting examined in most detail in Dell's (1986) original modelling endeavours was the number of timesteps per selection stage. As the implementations presented by Dell (1986) focused entirely on phonological encoding with output at a phoneme layer, this was equivalent to the number of times activation was calculated at each representation after a target morpheme had been jolted and before the most activated phoneme was selected.

Dell (1986) suggested that there was a direct correspondence between the number of steps before selection and speech rate in humans, such that fewer steps represented a quicker speech rate, and more steps represented a slower speech rate. Previous empirical studies (MacKay, 1971) and experimental investigations by Dell (1986) show that humans make more errors at faster speech rates. Dell (1986) argued that this should also be the case in the spreading activation model. With fewer timesteps before phoneme selection, Dell (1986) reasoned that target phonemes may not receive enough activation from the higher level morphemes to compete successfully, and that old items may still be activated as they have not had enough time to decay.

Dell (1986) reports on a simulation employing a network in which there are one or two nodes representing syllables and their structure (e.g., consonant clusters and rimes) between the morpheme and the phoneme layer. Results are reported for productions of chains of one, two or six randomly selected two syllable words, with two, three, four or five steps between morpheme jolt and phoneme selection.

This simulation appears to confirm that the model is more erroneous when there are fewer timesteps before selection. However, when evaluating these results, it should be borne in mind that there is no activation-based or intrinsic noise in this simulation. Any errors which occur are entirely due to either a lower number of timesteps than intermediary nodes between the morpheme and phoneme nodes, or because production of previous words has left some nodes in the network very activated, such that new target nodes cannot compete[2]. Examination of the error rates shows that the timestep effect is largely driven by the two timesteps condition, in which around 60% of phoneme productions are erroneous. For phonemes which are separated from the jolted morpheme by a syllable node and either rime or cluster node (i.e., all the vowel and coda phonemes, and any onset phonemes which participate in an onset cluster), the activation transmitted by the morpheme will only reach the rime or cluster node, and absolutely no activation will reach the phoneme node until the following timestep. Phoneme selection is in these cases completely random. For productions of two word or six word strings, perseveration of activation is indeed a greater problem with fewer timesteps and causes slightly higher error rates than when more timesteps are allowed, because as Dell (1986) explained, in with fewer timesteps, activation has less time to decay. However, there are no errors at all for single word productions with three, four, or five steps before selection, because neither of these two problems can apply.

Dell (1986) reports another investigation, simulating an experimental error elicitation paradigm known as the SLIP task (e.g., Baars et al., 1975). In this task, participants are asked to produce two word phrases, such as *"mad back"*. However, before being cued to produce the target phrase, they are shown a series of phrases in which the onsets are reversed, such as *"bid meek"*, *"bud muck"* and *"big men"*, to prime the reversed onsets and increase the probability of an onset exchange on production of the target phrase (e.g., *"mad back"* → *"bad mack"*). Simulation of this task using two, three or four steps before selection also appears to show that more errors are generated when fewer timesteps are allowed. Again however, there is no activation-based or intrinsic noise in this simulation. Errors are caused by random amounts of activation being added to the onset of the second word at the start of the encoding of the first word (*anticipatory bias*), and to the onset of the first word

[2]A further potential source of errors would be activation growing without bound due to an inappropriate combination of connection weight and decay, as discussed earlier, but Dell (1986) chose parameters (connection weight = 0.3, decay rate = 0.6) which avoid this situation and meet the constraint set by Shrager et al. (1987), such that the connection weight is lower than the decay rate.

at the start of the encoding of the second word (*perseveratory bias*), simulating the reversed onset bias in the SLIP task.

None of the errors in this simulation are due to activation not reaching the target phoneme on time, as morphemes are directly connected to phonemes. However, this simulation does further illuminate the effect of timesteps on perseveratory activation and therefore the production of movement errors, particularly exchanges. Anticipations and perseverations are not greatly affected in this simulation, as their generation is almost entirely determined by the amount of anticipatory or perseveratory bias applied (and there are very few neighbours to cause perseveratory productions of onsets in the normal way). A lower number of timesteps produces many more exchanges however, as there is less time for the activation of the intended but unselected first onset to decay. The number of exchange errors produced has a very big effect on overall error rate in this simulation as with the support of the anticipatory and perseveratory bias, this model produces a very high proportion of exchanges, with many more exchanges than anticipations or perseverations. This is not the behaviour of Dell's (1986) standard model which aims to capture corpus data patterns, but is in line with the experimental SLIP task human performance data collected by Dell (1986). The more prolific generation of exchange errors with fewer timesteps also matches up with Dell's (1986) empirical SLIP evidence, which shows more exchange errors at faster speech rates.

The main phonological encoding simulation reported by Dell (1986) does include activation noise, with a standard deviation of 0.2. In this simulation, the model produced random two word phrases with either three, four or eight steps before phoneme selection. The architecture in which there were syllable nodes and rime and cluster nodes between the morpheme and phoneme nodes were employed again. This simulation gives some suggestion that more errors are caused when fewer timesteps are allowed, but the results are less clear cut. Analysis showed that 153 phoneme errors were recorded with three timesteps and 72 phoneme errors with four timesteps. Perseveration errors comprised a large component of the increase in errors at three timesteps. This follows since perseverations are generated by neighbour reactivation in this simulation, and neighbour activation will decay over time. Exchange errors also make some contribution, again because with fewer timesteps, there is less time for the activation of the unselected intended first onset to decay. However, 79 errors were recorded at eight timesteps, which was not very different to the number of errors generated at four timesteps.

Dell (1986) also notes that when more timesteps are allowed before selection, interactive effects become stronger, because there is more time for activation to reverberate in the network. Specifically, Dell (1986) argues that the lexical bias effect should be stronger when there are more timesteps before selection, and his simulations with his SLIP network confirm this prediction. Both the number of word outcome and the number of nonword outcome errors decrease as more timesteps are allowed before selection, but the number of nonword outcome errors decreases more rapidly, as unlike word outcomes, nonword outcomes are not supported by a morpheme node. This predicted increase in lexical bias effects as speech rate decreases is supported by Dell's (1986) empirical results. Similarly, Dell (1986) shows that in the simulation, the number of features shared by phonemes involved in an error increases with the number of steps before selection. Again, a higher number of steps before selection makes more time available for feedback loops between features and phonemes to affect phoneme activation levels.

The number of lexical outcome errors is further affected by whether the word outcome shares a vowel with the intended production or not. Empirical results have shown that people make more errors when more phonemes are shared, an effect known as the *repeated phoneme effect* (e.g., Dell, 1986; MacKay, 1970). Dell's (1986) model explains this effect by reference to feedback, as more activation can pass to the competing lexical representation if they share more representations at the phonological level. In a similar pattern of results to the lexical bias results, both word outcome errors which do and don't share a vowel with the intended production decrease as more timesteps are allowed before selection, but word outcome errors which do share a vowel decrease less rapidly as they are better supported. However, it should be noted that in Dell's (1986) empirical results, there is only a numerical trend rather than a statistically significant effect of speech rate on the repeated phoneme effect.

A rather different effect of the number of steps is alluded to by Goldrick and Rapp (2002) however. Goldrick and Rapp (2002) suggest that damage to a level of representations in their model could partially be simulated by assuming that selection occurs after a greater number of timesteps. This would permit more activation to cascade from the damaged level, and more activation to feed back. For example, if selection took longer at the lexical level, more activation could cascade to phonemes, which would then support feedback of activation to formal neighbours at the lexical level. Whilst their simulations do indeed show that more formal errors occur in this

scenario, the jolt is decreased and noise at the lexical level is increased at the same time, and so it is not possible to see the individual effect of the steps manipulation.

To summarise this long examination of the effect of manipulating the steps parameter, Dell's (1986) original theory made an inverse connection between the number of steps per selection stage and speech rate. His investigations suggested that more errors are made when there are fewer timesteps before selection, that more perseveration and exchange errors are generated with fewer steps, but that interactive effects are stronger when there are more steps before selection. However, Goldrick and Rapp's (2002) investigations make a small challenge to this first assertion that fewer errors occur when more steps are allowed, suggesting that in fact more errors occur. Notably, in this final model, there is some noise on representations, whereas in the simulations which strongly suggest that fewer errors occur as more timesteps are allowed, no noise is present (Dell, 1986).

Activation-based noise

The most detailed examination of manipulating the activation-based noise parameter has been carried out by Rapp and Goldrick (2000; Goldrick & Rapp, 2002; Goldrick, 2006; see also Ruml et al., 2000, 2005) with the aim of simulating aphasic damage at various levels in the production system. The results of these simulations have largely been reported in the previous sections as other parameters have often been manipulated alongside. The general result however is that noise at any level of representation unsurprisingly leads to more errors in selection at that level. If the jolt size is low, noise from a damaged level can cascade and disrupt processing at a lower level. If the feedback connection strength is sufficiently high, damage caused by activation-based noise can lead to lower levels having an increased effect on selection at a given level due to feedback, but also to higher levels being disrupted due to noise from the damaged level feeding back.

Intrinsic noise

In contrast to activation-based noise, very few investigations of manipulation of the intrinsic noise parameter have been reported. As noted previously, Dell, Schwartz, et al. (1997) do make a theoretical argument that the intrinsic noise parameter has the inverse effect of the jolt parameter in single word production models however, suggesting that the jolt parameter simply sets the scale of activation, and doubling the intrinsic noise parameter would have the same effect as halving the jolt parameter. Furthermore, as also noted above, they present simulations suggesting that

increasing the intrinsic noise parameter has a similar effect to increasing the decay parameter whilst slightly reducing connection strength (a suggestion echoed by Rapp & Goldrick, 2000). For simplicity, they focus on manipulations of the decay parameter when trying to simulate aphasic evidence, but claim that they could have manipulated the intrinsic noise parameter to similar effect.

2.4.2 Investigating architectural options within the spreading activation model

The evidence summarised above demonstrates the many ways in which manipulating the parameters of the spreading activation model strongly affects its behaviour. In our attempt to investigate which phonological encoding to subphonemic process information flow options can account for new VOT, EPG and ultrasound evidence (Goldrick & Blumstein, 2006; McMillan, 2008; McMillan et al., 2009), we are proposing different mechanisms to account for the behaviour patterns in different information flow architectures. The previous section makes it clear that certain mechanisms work better under different parameter settings; for example, perseveratory mechanisms are more effective with fewer timesteps per selection stage, and interactive mechanisms are more effective with more timesteps per selection stage. Unless we can theoretically rule out the possibility that the different accounts that we have proposed may also benefit from different parameter settings, we need to pay careful consideration to how we deal with the parameters when comparing these different architectures.

Only a few other researchers have tried to compare architectural options in the spreading activation model. A number of these have focused on models in which a small number of parameters (at most two) are manipulated to allow the model to fit individual aphasic patient error patterns for a group of patients (Dell et al., 2004; Foygel & Dell, 2000; Ruml et al., 2000, 2005). As at most two parameters are manipulated in any one model, it is not entirely clear what the influence of other parameters is on the models' capabilities to account for this data. Furthermore, this per-patient data is clearly rather different in nature to the qualitative statistical observations about the normal population which we hope our models will be able to account for.

Other investigations consider data more similar in nature to ours. Hartsuiker (2002) investigates the ability of two models to account for the tendency of normal speakers to add rather than delete phonemes from a word when making a phonological error, the *addition bias* (Nooteboom, 1969). For example, a speaker is more likely to

produce a CVC syllable in place of a CV syllable than vice versa. The two models he investigates are based on Dell (1986) and Dell (1988). These two models differ in the way in which they produce syllables of different shapes. The model based on Dell (1986) has one frame shape (CVC), and smaller syllables such as CV syllables are produced by filling the empty slots with null segments. The model based on Dell (1988) has many frame shapes, and does not include null segments. Both models are implemented using representations at a syllable layer which are connected to syllable shapes as well as phonemes. A syllable shape is selected at the same time as the phonemes, and this determines the shape of the frame in which the phonemes must be placed and hence the shape of the produced syllable. In the model based on Dell (1986), there is only one syllable shape node, and there are null phonemes available for selection. In the model based on Dell (1988), there are multiple syllable shape nodes, and no null phonemes.

Hartsuiker (2002) found that the model with multiple syllable shape nodes predicts an addition bias, whereas the model with one syllable shape node does not; in fact, as Dell (1986) reported, it predicts a deletion bias, which is not in line with the empirical evidence. The model with multiple syllable shapes accounts for the addition bias as syllables with more phonemes (e.g., CVC syllables) receive more activation via feedback from phonemes to syllables than syllables with fewer phonemes (e.g., CV syllables). Bigger syllables then transmit more activation to both their syllable shape and their component phonemes. In contrast, this argument does not work in the model with a single syllable shape, as all syllables are connected to three phonemes, possibly including a null phoneme. Furthermore, the null phonemes are very frequent phonemes as they appear in every non-CVC syllable, and as such, feedback between the syllable layer and the phoneme layer renders them very active and very likely to be selected, creating a deletion error. These results are further developed with a simulation investigation of the role of the contents of the lexicon, which shows that addition bias is stronger in languages such as Dutch where bigger syllables are more common than smaller syllables, than in languages such as Spanish where bigger syllables are less common, a finding in line with corpus analyses presented by Hartsuiker (2002). Furthermore, Hartsuiker (2002) shows that stronger feedback between phonemes and syllables increases the addition bias, and that feedback between syllable shapes and syllables increases the addition bias in languages where bigger syllables are more common, but decreases it in languages where smaller syllables are more common (see also section 2.4.1).

This investigation of model parameters, such as the feedback connection strength, is useful and interesting, but not much consideration is given to the role of other spreading activation parameters in this result. Admittedly the reasoning given above makes it seem very unlikely that Dell's (1986) architecture would work at other parameter settings, but without further exploration of the parameter space, it needs to be borne in mind that this is still an assumption made based on pen and paper reasoning. More importantly however, is the success of the Dell (1988) model entirely due to the architecture, or do the parameters chosen also have a role to play? In other words, would this model work at any parameter setting? If not, why not?

A final approach to comparing different architectures in the spreading activation model is presented by Rapp and Goldrick (2000). Rapp and Goldrick (2000) compare models of information flow between semantic and lexical representations, and lexical and phonological representations, and evaluate their ability to account for multiple qualitative patterns of evidence, such as the lexical bias effect, the mixed error effect, and the absence or coexistence of certain error types in a small set of aphasic patients. Both the qualitative nature of the behavioural patterns, and the theoretical focus on information flow in this study bear strong resemblance to the empirical results and questions we wish to deal with in this thesis. Concentrating on finding models which can account for the broad qualitative patterns in empirical data also seems like an inherently sensible approach when attempts to model word production are at such an early stage. Furthermore, Rapp and Goldrick (2000) explicitly consider the role of feedback connection strength, jolt size, and to the extent that it is used to model aphasic damage, activation-based noise, on the models' ability to account for the data. However, there is still some room to make improvements upon this approach. Firstly, where Rapp and Goldrick (2000) rule whole information flow architectures out, they do this based on their behaviour at one parameter setting. As in Hartsuiker (2002), the arguments presented as to why the models would not be able to account for this data at any setting are fairly convincing. However, one of the major advantages of evaluating theories via computational simulation is the opportunity to verify that no aspect of theory behaviour has been missed by a pen and paper approach. Not checking behaviour at multiple settings means that the power of the computational approach is not fully leveraged. Secondly, Rapp and Goldrick (2000) do not vary all parameters when investigating a model's ability to account for the data. Whilst they comment on the potential effect of manipulating other parameters such as decay and intrinsic noise, it would be preferable to back up such theoretical suggestions with simulation data.

2.4.3 Summary

This section has focused on the spreading activation parameters which determine the way in which activation flows around a spreading activation model such as that proposed by Dell (1986). These are forward connection strength, feedback connection strength, jolt size, prime, decay, number of steps, activation-based noise, and intrinsic noise. It was shown that there is no clear method by which settings for these parameters can be derived directly from human characteristics. Previous studies have used various different settings, as summarised in table 2.4, and therefore also do not dictate what should be chosen.

The effects of manipulating various parameters, as revealed in previous studies, were summarised. Studies manipulating connection weight have suggested that too low a *connection weight* (for both forward and feedback connections) can lead to inconsistencies in activation patterns at different levels, a reduction in interactive effects, and an increase in errors overall (Dell & Gordon, 2003; Dell, Schwartz, et al., 1997; Foygel & Dell, 2000; Ruml et al., 2000, 2005). Other studies suggest that the connection weight must be lower than the decay rate for the network to behave reasonably (Shrager et al., 1987). *Feedback connection strength* manipulations showed that higher feedback strength can lead to better reinforcement of intended productions (Ruml et al., 2000) and makes production of both bigger and more frequent representations more likely (Hartsuiker, 2002). However, if feedback is too strong, particularly to or from a damaged process, higher level processes can be unduly disturbed by lower level processes (Goldrick, 2006; Rapp & Goldrick, 2000). *Jolt size* serves as both an index of selection strength (Goldrick, 2006; Rapp & Goldrick, 2000), as well as determining the amount of activation in the network overall, which affects the influence that other parameters such as intrinsic noise can have on the behaviour of the network (Dell, Schwartz, et al., 1997). The effect of *prime* size has not really been investigated, although Dell (1986) highlights that a higher prime should lead to more anticipations and exchanges being produced. In investigations of single word production, it has been shown that a high level of *decay* leads to the model making many errors, although most errors tend to maintain some relation to the target (Dell & Gordon, 2003; Dell, Schwartz, et al., 1997; Ruml et al., 2000). Investigations of multiple word production (Dell, 1986) have suggested that lower decay levels lead to increased perseveration and exchange rates, where exchanges are produced at the expense of anticipations. If decay is reduced too far, strings of nonsensical perseverations ensue. Dell (1986) also suggested that fewer *steps* per selection stage models a faster speech rate, such that more errors should

occur when there are fewer timesteps before selection. Simulations which strongly confirm this suggestion do not have any noise on representations however. There is some suggestion from other investigations that if there is noise in the network, a higher number of steps may cause more errors (Goldrick & Rapp, 2002). Dell (1986) also suggests that more perseveration and exchange errors are generated with fewer steps, but that interactive effects are stronger when there are more steps before selection. Finally, *activation-based noise* has been used to simulate aphasic damage, causing more errors at the noisy level (Goldrick & Rapp, 2002; Goldrick, 2006; Rapp & Goldrick, 2000; see also Ruml et al., 2000, 2005). Rapp and Goldrick (2000; Goldrick & Rapp, 2002; Goldrick, 2006) have also shown that a low jolt size can lead to noise at a damaged level affecting processing below, and that a high feedback strength can lead to damaged processes disturbing processes above, or being disturbed by processes below. *Intrinsic noise* manipulations have not been reported in depth, although researchers have suggested that high intrinsic noise will lead to the same behaviour as in a network with high decay (Dell, Schwartz, et al., 1997; Rapp & Goldrick, 2000) and that halving intrinsic noise is the same as doubling jolt size (Dell, Schwartz, et al., 1997).

Clearly then, some mechanisms generate stronger effects at some parameter settings than others. It was therefore suggested that when comparing the different accounts of phonological encoding to subphonemic process information flow proposed in section 2.3.2, careful consideration will need to be paid to how we deal with the parameters. Approaches taken by other researchers towards comparing architectures in the spreading activation model were summarised. A number of authors have focused on modelling individual differences in aphasic behaviour, dealing with per patient error patterns (Dell et al., 2004; Foygel & Dell, 2000; Ruml et al., 2000, 2005). Whilst these models all depend on the manipulation of up to two parameters to capture the individual differences in aphasic data, the role of the other parameters in determining the model's ability to fit the data is not demonstrated. Furthermore, the models are evaluated using quantitative comparisons of the model's predictions and patient data, whereas our data comprises statistically supported qualitative observations about the normal population. Hartsuiker (2002) reported on a comparison of the ability of two models to account for a single qualitative pattern, similar to the patterns we are interested in in this investigation. The contribution of feedback strength to the successful model's ability to account for the data is examined, but there would perhaps be more room to clarify the role of the other seven activation spreading parameters in the model's success. Finally,

Rapp and Goldrick (2000) compare the ability of a number of models with different semantic to lexical and lexical to phonological information flow assumptions to account for multiple qualitative patterns. Both the shape of the data and the theoretical focus closely resemble the data and questions that this thesis is concerned with. Furthermore, Rapp and Goldrick (2000) demonstrate the effect of manipulating feedback connection strength, jolt size and activation-based noise on the ability of the different models to account for the data. However, it was suggested that it would be possible to further develop this approach by leveraging computational power to more clearly demonstrate whether certain accounts work by testing them at other parameter settings, and to explore the effect of manipulating all of the parameters, rather than relying purely on theoretical prediction.

2.5 Chapter summary

The first section of this literature review focused on questions that speech error models have traditionally addressed. How do units become misordered in word production? And how does information flow between the two hypothesised stages of lexical selection and phonological encoding in the two-stage model of word production? It was found that the theory proposed by Dell (1986) is the only one which has led to implementations which successfully simulate both movement errors and account for all the speech error evidence considered in the information flow debate, by implementing cascading from lexical selection to phonological encoding, and feedback from phonological encoding to lexical selection. It appears that there may be some problems with comparisons between the behaviour of Dell's (1986) model and corpus evidence however. A reexamination of the corpus data and the behaviour of Dell's (1986) model in this respect is presented in chapter 5.

The second section considered the assumption in Dell's (1986) model that subphonemic errors do not occur, and that the only information transmitted from phonological encoding is the identity of the selected phonemes. It was argued that the transcription data upon which these claims are based does not appear very reliable in the light of results from the perceptual literature and recent instrumental investigations of speech production. The section then considered three experiments using instrumental measurements of speech output (Goldrick & Blumstein, 2006; McMillan, 2008; McMillan et al., 2009). Four models of information flow were presented: no cascading from phonemes, cascading from selected phonemes only, cascading from all phonemes, and feedback from subphonemic representations. It was argued that contrary to Goldrick and Blumstein's (2006) claims, cascading from

all phonemes including unselected phonemes is not needed to account for their data, and that instead, models with no cascading from phonemes can account for the large part of the data, and models with cascading from selected phonemes only can account for all of it. All models were hypothesised to be able to account for the results presented by McMillan et al. (2009). However, as suggested by McMillan (2008), it is expected that feedback from subphonemic representations will be required to account for his instrumental measures of manipulations of phonological similarity. It was argued that this is particularly interesting, as the transcribed phonological similarity effect will no longer provide evidence for feedback from subphonemic representations if it is accepted that subphonemic errors can occur. Simulations of the transcribed phonological similarity effect without this assumption are presented in chapter 7. Simulations of the data presented by Goldrick and Blumstein (2006), McMillan (2008) and McMillan et al. (2009), which will also constitute the first simulations of instrumental data using a model based on Dell (1986), are then presented in chapters 7 and 8.

The third and final section considered the problem of comparing information flow assumptions when there are so many other free parameters in the spreading activation model. The section summarised findings from the previous literature which throw some light on the potential effects of manipulating these parameters, and also described previous approaches to comparing architectural options in the spreading activation model. Examination of this previous work highlighted a need for an approach which allows us to investigate the ability of a model to account for multiple qualitative patterns. In cases where an architecture cannot account for the data, the power of the simulation approach should be leveraged to demonstrate that the problem truly lies with the architecture, not just the parameter settings chosen. In cases where an architecture can account for the data, the role of the parameter settings in this result should be clarified. Development of new methodology meeting these requirements and exploration of the effect of manipulating parameters on basic measures such as error rate and contextuality of errors, and more complex error behaviours such as directionality of movement errors and lexical bias and phonological similarity, is presented in chapters 3, 4, 5, and 6.

Chapter 3

Model implementation

3.1 Introduction

Chapter 2 presented a number of hypotheses about the ability of different models of information flow between phonological encoding and subphonemic processing to account for new instrumental data (Goldrick & Blumstein, 2006; McMillan, 2008; McMillan et al., 2009). In order to test these hypotheses, it was necessary to implement an extension of Dell's (1986) model. This implementation permitted output at the subphonemic level, and a method was created to permit comparison of the output to acoustic measures such as VOT, and articulatory measures such as EPG and ultrasound.

In chapter 2, the issue of parameter settings in the spreading activation model was also raised. A need for awareness of the effect of parameter settings on model behaviour was highlighted, and in particular, it was argued that there would be problems with testing the ability of different models of information flow to account for the data at one arbitrarily chosen parameter setting. Firstly, if one information flow model was found not to work, the simulation evidence would not show whether this was due to a general inability of the information flow model to account for the data, or whether it was rather the combination of the information flow model and that particular set of parameters. Secondly, if it was found that an information flow model did work, it would not be clear to what extent this result was really dependent on the parameters chosen, rather than simply the choice of information flow model.

This chapter outlines the basic implementation of the models, which followed principles laid out by Dell (1986). The implementation was tested with output at the phoneme level and output at a subphonemic level, in order to highlight some of the

effects of changing the output level. In this chapter, the word production procedure in both of these modes is explained, alongside the interpretation of the output for simulation of both transcribed and instrumental evidence. The chapter also describes the implementation of the various information flow options. This includes the four options for information flow between phonological encoding and subphonemic processes. In addition, feedback between phonological encoding and lexical selection was also manipulated to demonstrate its effects on the models' ability to account for the various pieces of transcribed and instrumental evidence. Finally, the issue of spreading activation parameter settings is addressed. The general approach of this thesis to variation of these settings is outlined, and the relationship between parameters used and those used in the previous literature is explained.

3.2 Representations in the model

Our implementation had three levels of representation: a word level, a phoneme level and a feature level.

A core focus of this thesis is the information flow between phonological encoding and subphonemic processes. The goal of simulations investigating this issue was to model evidence demonstrating lexical and phonological influences on error patterns at a subphonemic level. Semantic effects were not considered at this point. For simplification therefore, words are directly activated in this implementation, and lexical selection via activation from semantic features is not currently simulated. Later work could add this stage to the implementation, however.

Subphonemic processes were modelled using a featural layer. This work does not intend to present a strong argument for a featural subphonemic representation, and later work may wish to consider other possibilities for representation at this level, such as gestures (e.g. Browman & Goldstein, 1989). However, features were the subphonemic representation used in Dell's (1986) original model, and so in the name of making incremental changes, the current study investigated what evidence can be accounted for without modifying the original representational assumption. Consonants were decomposed into place, manner and voicing features, and vowels were decomposed into height, backness, and roundedness features. Activation levels of the onset consonant voicing features were used to simulate VOT evidence, and activation levels of the onset consonant alveolar and velar place features were used to simulate EPG and ultrasound evidence. Interpretation of this output is explained in more detail in section 3.4.2.

Between the word and the feature representations was a phoneme layer. This implementation focused on the production of monosyllabic CVC nodes for simplicity, and therefore, as in many models before (Dell, 1986, SLIP task model; Dell, 1990; Dell et al., 2004; Dell & O'Seaghdha, 1991, 1992; Dell, Schwartz, et al., 1997; Foygel & Dell, 2000; Goldrick, 2006; Goldrick & Rapp, 2002; Martin et al., 1994; Oppenheim & Dell, 2008; Rapp & Goldrick, 2000; Ruml et al., 2000, 2005; Schwartz et al., 2006) words were directly connected to phonemes, and there were no syllable nodes or consonant cluster and rime nodes in between, unlike in the first model presented by Dell (1986). As all words were CVC words, no null phonemes were included (cp. Dell, 1986), and for simplicity, no wordshape nodes either (cp. Dell, 1988; Hartsuiker, 2002).

As in Dell's (1986) original model, both phonemes and features were grouped according to their syllable position: onset, nucleus or coda. Consonants and consonant features which occurred in both onset and coda position were represented in the network twice, once for each syllable position. For example, if the network vocabulary contained both the words *cap* and *back*, a node for the phoneme /k/ would be included in the onset phoneme group, and another node for the onset /k/ would be included in the coda phoneme group. Similarly, at the featural level, the consonant features *velar*, *stop* and *voiceless* would be represented by separate nodes in both the onset and the coda syllable positions. Features were also grouped according to their type, with a place, manner and voicing feature node group for the onset and coda positions, and a height, backness and roundedness group for the nucleus position.

The lexicon of the network was determined separately for each experiment. In each case, a vocabulary of either 50 or 100 words was created. Methodology for vocabulary generation is explained alongside the simulations in later chapters. Once the word nodes for each experiment had been determined, the phonemes necessary to produce the words and the features necessary to produce the phonemes were added to the network.

A mini network suitable for encoding the words gap and cap is presented in figure 3.1.

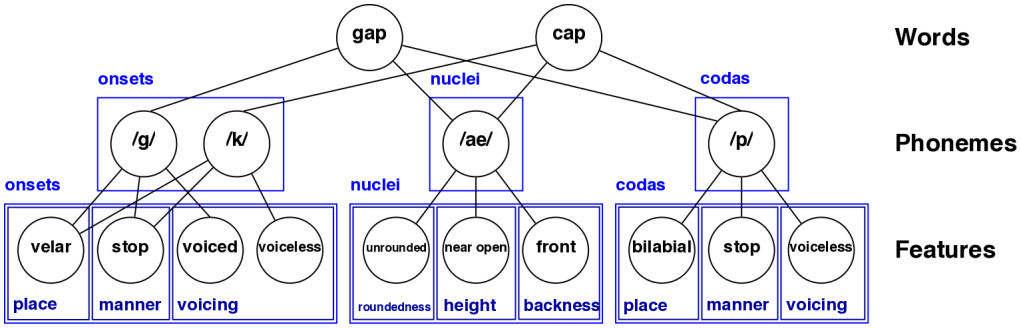


Figure 3.1: A mini network suitable for encoding the words gap and cap.

3.3 Processing stages in the model

Whilst Dell's (1986) original model assumed output at the phoneme level, it was argued in the previous chapter that the evidence that this assumption was based on was not convincing. Furthermore, to simulate the instrumental data we are interested in, output at a subphonemic layer is required. However, it was expected that changing this assumption would mean that feedback would no longer be required from the subphonemic layer to the phoneme layer in order to account for the transcribed phonological similarity effect. A change in output level could also have an effect on other basic behaviours of the implementation, such as the error rate. To demonstrate the effect of making this change, two processing modes were implemented: a *phonological encoding only* mode, with output at the phoneme level, and a *phonological encoding and subphonemic processing* mode, with output at a subphonemic level. These two modes are described below.

3.3.1 Phonological encoding only

In the *phonological encoding only* mode, the implementation behaves like Dell's (1986) original model. Production of a single word begins with a jolt to the target word, and ends after the specified number of timesteps with selection of the most activated phoneme in each syllable position group at the phoneme layer. If there is feedback from the subphonemic layer to the phoneme layer, then featural activation can affect phonological encoding, but features are not selected or considered as output.

3.3.2 Phonological encoding and subphonemic processing

In the *phonological encoding and subphonemic processing* mode, an extra stage is added following phonological encoding. At the end of phonological encoding, the

most activated phoneme in each syllable position group is both selected, and given a large jolt of activation[1]. Activation then spreads again for the specified number of timesteps, after which output is read from the subphonemic layer.

The following section explains how output is interpreted in both of the two processing modes.

3.4 Model output

3.4.1 Simulating transcribed evidence

Transcriptions of speech errors are typically represented phonemically: for example, a [k] produced when a /g/ is intended. To simulate transcribed evidence therefore, some assumptions need to be made about how the listener would interpret the output of the model.

In the phonological encoding only mode, it was assumed that the listener would hear and transcribe the phonemes selected at the end of phonological encoding. This assumption was also made in Dell's (1986) original model and all other extensions of this model which have been built so far.

Much evidence suggests that listeners tend to categorise speech-like sounds as phonemes (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Accordingly, in the phonological encoding and subphonemic processing mode, output was classified as the phoneme formed from the most activated features in each syllable position at the end of subphonemic processing. Feature combinations which did not form a phoneme in the English phoneme inventory were recorded but were not assigned a phonemic category.

3.4.2 Simulating instrumental evidence

Instrumental approaches allow us to measure speech production directly, without potential categorical bias caused by the human perception system. To simulate instrumental measurements, specifically VOT, EPG and ultrasound, the activation levels of onset consonant features were recorded.

[1]Under most information flow assumptions, the jolt of activation is added to the phoneme's current activation, but under others the phoneme's activation is set to the jolt amount. See section 3.5 for further details.

Voicing onset time

Voicing onset time is a one dimensional measure of consonant production representing the (positive or negative) time between the release of the relevant articulators, and the onset of vocal cord vibration. Here, VOT was simulated as the activation of the voiceless feature minus that of the voiced feature. This measure varies in a similar way to voicing onset time, such that very voiceless productions result in a high value, and voiced productions result in a low value.

EPG and ultrasound

EPG and ultrasound provide measures of the movement of articulators, particularly the tongue. To allow for a simulation of EPG and ultrasound measures, all targets in this thesis had alveolar or velar onsets, in line with previous instrumental investigations (e.g. McMillan, 2008; McMillan et al., 2009). Here, we took the degree of activation of the alveolar and velar features to abstractly represent the extent to which the resulting articulation involves tongue raising at the front and the back of the mouth respectively. For example, a velar stop such as /k/ would normally be associated with high velar and low alveolar activation. However, a high alveolar activation alongside the high velar activation would represent a /t/-like influence or intrusion on the production. In these simulations, only one measurement of featural activation is made, at the end of subphonemic processing. Timing of tongue movement is therefore not explicitly simulated.

Analysis by the delta method

Productions measured using EPG, ultrasound, or VOT can be compared using the delta method, described in section 2.3.2. This method can equally be applied to both vectors of alveolar and velar feature activation, and measurements of VOT. For alveolar and velar vectors, the delta method can be implemented as the Euclidean distance between two vectors, and for measurements of VOT, delta is equivalent to the absolute difference between the two readings.

3.4.3 Focus on onset productions

It has been shown that phonemes in the onset position are much more prone to error than phonemes elsewhere in a syllable. For example, MacKay's (1970) analysis of Meringer's speech corpora (Meringer & Mayer, 1895; Meringer, 1908) demonstrated that 81% of between word consonant exchanges, and 96% of within word consonant

exchanges occur at the start of the syllable. Vowel exchanges also occur about as infrequently as final consonant exchanges, according to MacKay's (1970) results.

Experimental speech error work tends to focus on onset errors because they are easy to elicit and also because such an approach should help maximise ecological validity of experimental results. Materials are created for these studies based on the assumption that errors will occur at the beginnings of syllables (e.g. Baars et al., 1975; Hartsuiker et al., 2005). Recent articulatory studies of speech errors, including those modelled in this thesis, have also chosen onset consonants as their target when measuring the phonetic characteristics of productions (e.g., Goldrick & Blumstein, 2006; McMillan, 2008; McMillan et al., 2009)

However, a remaining weakness of Dell's (1986) model is its inability to replicate the human tendency to make more errors at the beginning of a syllable. In Dell's (1986) model, all phonemes are encoded simultaneously, and no theoretical distinction exists between onset phonemes and phonemes later in the syllable. Correspondingly, Dell's (1986) simulations showed that the model is just as likely to produce errors on these later phonemes as error at the syllable onset.

Clearly, as highlighted by Dell (1986, 1988), work is required to make the model much less likely to produce errors in the rime of a word than in the onset. Pending such improvement however, error status classifications in the present thesis focused solely on onset outcomes. For experiments in which materials were manipulated, for example to contrast lexicality of error outcome, these manipulations were on the onset. Post-onset errors were not interpreted as these materials were not designed to elicit such errors, but materials were controlled such that the rimes matched. To facilitate comparison with these experiments, output from simulations based on random word production was analysed at the onset only.

3.4.4 Zero selections

During preliminary simulations to test the production mechanism of our implementation we noticed that on rare occasions, the selection process would find that all the nodes in a group had activation levels of zero. The assumption was adopted that in this situation, from here on referred to as a *zero selection*, the human production system would abandon the utterance. Productions where this situation was encountered were therefore aborted and excluded from analyses. In the next chapter, we demonstrate that in most specific models we tested, no productions were aborted

for this reason. We also describe the characteristics of the specific models which were affected by this problem, albeit on a very low percentage of productions.

3.5 Information flow

In section 3.3, we detailed modifications to the processing stages in Dell's (1986) original model. Specifically, we implemented two versions of the model: a one-stage phonological encoding model with output at the phoneme level, modelled closely on Dell (1986), and a two-stage model which added subphonemic processing and output. In the present section, we describe implementations which modify the flow of information between levels of representation proposed by Dell (1986).

In Dell's (1986) original model, the same model of activation flow was assumed between all levels of representation. Firstly, at each timestep, activation flowed from all nodes in a level to all connected nodes at the subsequent level. This flow of activation from all nodes at all times is referred to as *cascading*. Secondly, at each timestep, activation also flowed back from all nodes in a level to all nodes in the previous level. This upward flow of activation is known as *feedback*.

The present thesis focuses largely on information flow between phonological and subphonemic representations. In the one-stage phonological encoding model based on Dell's (1986) original implementation, we tested the model both with *feedback from subphonemic representations*, and with *no feedback from subphonemic representations*. This manipulation was primarily intended to demonstrate that feedback from subphonemic representations is required for the model to exhibit a phonological similarity effect when output is at a phoneme level.

The key predictions laid out in chapter 2 concerned information flow between phonological encoding and subphonemic processing in the two-stage model, where output at a subphonemic level makes simulation of instrumental evidence possible. Thus a particular focus of the thesis is the four possible information flow models discussed earlier, and presented in table 2.2, replicated in table 3.1 for convenience.

The most discrete model assumes that only the identity of the selected phoneme is conveyed to subphonemic processes. At the end of phonological encoding, the activation of the most active phoneme is set to a pre-specified jolt amount. This is the only activation which is transmitted from phonemes to the featural level. No activation is conveyed from phonemes to features prior to selection at the phoneme level. We describe this model as having *no cascading from phonemes*.

Table 3.1: Activation flow characteristics of the four proposed models of information flow between phonological encoding and subphonemic processes (replicated from table 2.2)

Model	**Information from phonological encoding**			**Feedback from subphonemic representations**
	Identity of selected phoneme	*Activation from selected phoneme*	*Activation from unselected phonemes*	
No casc	✓			
Casc from sel	✓	✓		
Casc from all	✓	✓	✓	
Feedback	✓	✓	✓	✓

The next model increases interactivity by allowing *cascading from selected phonemes only*. In this model, the activation of the most active phoneme at the end of phonological encoding is incremented by the jolt amount. The activation transmitted to the featural level will therefore differ depending on how strongly activated the phoneme was prior to its selection. No activation is conveyed from phonemes to features prior to selection at the phoneme level.

The third model allows *cascading from all phonemes*, whether or not they have been selected. Since selection is not a pre-condition of cascading, activation cascades from phonemes both before and after the most active phoneme is selected and its activation level is incremented by the jolt amount.

The final model is the most interactive model and includes *feedback from subphonemic representations*. In this model, activation cascades from phonemes and feeds back from subphonemic representations both before and after selection at the phonological encoding stage.

Previous multi-stage models (e.g., Dell, Schwartz, et al., 1997; Rapp & Goldrick, 2000) have implemented post-selection jolt by setting the activation level of the selected representation to the jolt amount, rather than incrementing the existing activation. This assumes that the degree to which a selected representation is activated prior to selection cannot influence lower levels in the model once selection is complete. To permit post-selection cascading from selected phonemes, all but the first of the previously described two-stage models increment rather than set the activation level of selected phonemes.

Lastly, we tested each one and two-stage model both with and without *feedback from phonemes* to the lexical level. In all cases, *cascading from all words* to phonemes was present. Although this manipulation was of secondary importance in this thesis, it allowed us to test our multiple parameter setting approach to architecture evaluation

using the well-established lexical bias effect, which should require feedback from phonemes to words. We also report the effect of phoneme to word feedback on other simulations of transcribed and instrumental evidence.

3.6 Spreading activation parameter settings

At each timestep t_i, activation of a node j is calculated using the following formula:

$$A(j,t_i) = max\{0, a(j,t_i) + G(a(j,t_i)s_{acti}) + G(s_{intrin})\}$$
$$a(j,t_i) = [A(j,t_{i-1}) + \sum_{k=1}^{n} p_{fwd}A(u_k,t_{i-1}) + \sum_{k=1}^{m} p_{fbk}A(d_k,t_{i-1})](1-q)$$

where

$A(j,t_i)$ is the activation level of node j at a particular timestep t_i;
$A(j,t_{i-1})$ is the activation level of node j at the previous timestep t_{i-1};
$u_1...u_n$ are the nodes with forward connections directly leading to node j;
$d_1...d_m$ are the nodes with feedback connections directly leading to node j;
p_{fwd} is the feedforward connection strength
p_{fbk} is the feedback connection strength
q, where $(0 < q < 1)$, is the decay rate;
$G(s)$ is noise, calculated by randomly generating a number from a Gaussian distribution with a mean of 0 and standard deviation s
s_{acti} is the factor by which the activation is multiplied to determine the standard deviation of the activation-based noise
s_{intrin} is the standard deviation of the intrinsic noise

Note that in the present model, we follow Dell (1986) in applying decay to the activation level calculated at the current timestep. Other authors (e.g. Dell, Schwartz, et al., 1997) have applied decay to the activation level calculated at the previous timestep. Intrinsic noise was added to the calculation in later models (Dell, Schwartz, et al., 1997) and did not feature in Dell's (1986) original formula. However, a lack of intrinsic noise can be emulated by setting s_{intrin} to 0.

As noted in section 2.4, eight parameter settings must be determined to calculate activation in the current model. The first five parameters are referenced in the

formula above. These are: *forward connection strength* (referred to as p_{fwd} above); *feedback connection strength* (p_{fbk}); *decay* (q); the factor by which the activation is multiplied to determine the standard deviation of the *activation-based noise* (s_{acti}); and standard deviation of the *intrinsic noise* (s_{intrin}). The remaining three parameters are: *jolt*, the amount of activation added to a selected representation; *prime*, the amount of activation added to the node representing an upcoming word; and *steps*, the number of timesteps at which activation is calculated per processing stage.

To leverage the power of computational simulations to improve our understanding of the operation of the proposed models, we propose to evaluate the behaviour of the models at multiple parameter settings. This will allow us to achieve two important goals. First, we will be able to make general claims about the behaviour of models with different combinations of processing stages and information flow options, while abstracting away from the effects of particular parameter settings. If a given model is found not to be able to account for the data, we wish to be able to show that this difficulty is independent of the parameter settings chosen. If a model is able to account for the data, we wish to be able to determine to what extent this ability is dependent on the parameter settings chosen. A second goal is to clarify in general the effects of particular parameter settings on the behaviour of spreading activation models.

The parameter values we used, shown in table 3.2, were based on values used in the previous literature, as summarised in table 2.4 in section 2.4. Specifically, the forward and feedback connection strengths chosen cover the range of strengths used in the previous literature, although some previous studies have used intermediate values not used here, as clear from table 2.4. The values chosen permit a range of ratios of feedback to feedforward strength to be tested. The jolt and prime values chosen permit a variety of selection strength and competitor influences to be tested, and cover the values used in many simulations. However, some previous simulations used very low jolt values between 1 and 10 (Dell, 1990; Dell & O'Seaghdha, 1991; Goldrick, 2006; Oppenheim & Dell, 2008; Rapp & Goldrick, 2000), and these are not covered by our range. The decay values tested here cover all of the decay values used in previous simulations, apart from the outlier of 0.2 in Dell (1990). The steps parameters chosen cover the range in the literature, with the exceptions of two simulations in which selection is not simulated (Dell, 1990; Dell & O'Seaghdha, 1991). Again, however, intermediate values have been used in previous studies which are not used here. Similarly, the range of activation-based noise factors are covered,

with the exception of the outlier value of 0.68 in Oppenheim and Dell (2008), where this value was not known to us at the time that parameter values were chosen. We note that a similarly high level of noise (0.7) is used to simulate aphasic damage in Goldrick (2006). Again, for activation-based noise, not all intermediate values are represented, and we do not test any models with no activation-based noise, as discussed below. Finally, two settings of intrinsic noise have previously been used in the literature. Here the range of possible values is extended to increase the potential for intrinsic noise to affect the models' behaviour.

In the present thesis, we aim to find specific models which occasionally make mistakes, at any word position. To achieve this goal, some randomness in the implementation's behaviour is required. We therefore do not test specific models with no noise affecting the activation calculation. A small number of previous models have not included any activation-based noise (see table 2.4). However, these models do not meet our requirements. The very first simulation described by Dell (1986) did not include any noise, but was incapable of generating certain words in certain word positions, due to insufficient time for encoding, or deterministic perseveratory activation in the network. The SLIP task simulation described by Dell (1986) relied on randomly applied anticipatory and perseveratory biases to generate errors, but we are looking for a model which generalises past one experimental task.[2] Finally, Dell and O'Seaghdha (1991, 1992) refer to a stochastic selection rule, but do not detail how this would be implemented.

Whilst most parameters are tested at three settings, jolt is tested at four settings as this parameter was thought to have particular potential to affect the simulations of Goldrick and Blumstein's (2006) evidence. The intrinsic noise parameter was also varied to four settings to allow a greater range.

All possible combinations of activation parameter values were tested over all information flow options, subject to two constraints. Firstly, the prime always had to be less than the jolt, so that the target word had at least some initial advantage over a competing primed word. Secondly, the feedback connection strength was never greater than the forward connection strength, as it was assumed that top-down activation should be at least as strongly conveyed as bottom-up activation. Nearly all previous models have observed this constraint (see table 2.4), with the exception of Rapp and Goldrick (2000) and Goldrick (2006), whose simulations show that the network behaves inappropriately when the feedback connection strength is stronger

[2]Note that the second and primary simulation of phonological encoding described by Dell (1986) did include activation-based noise.

Table 3.2: Activation parameter values used in simulations

Name	Short name	Values
Forward connection strength	*fwdConn*	0.05, 0.2, 0.35
Feedback connection strength	*fbkConn*	0.05, 0.2, 0.35
Jolt	*jolt*	50, 100, 150, 200
Prime	*prime*	10, 50, 100
Decay	*decay*	0.4, 0.5, 0.6
Steps	*steps*	2, 5, 8
Activation-based noise	*actiNoiseSD*	0.05, 0.15, 0.25
Intrinsic noise	*intrinNoiseSD*	0, 0.005, 0.01, 0.05

than the forward connection strength. Future work could further investigate specific models in which the feedback connection strength is stronger than the forward connection strength if so desired.

For architectures which included feedback connectivity, there were 5832 eligible combinations of parameters. For architectures without feedback, the feedback connection strength was not varied, resulting in 2916 eligible combinations.

The same parameter settings were used in every simulation. This approach facilitates comparison of models between different simulations, highlighting which parameter settings can account for multiple types of evidence.

3.7 Implementation details

Combining processing stage options, information flow options and parameter settings resulted in a large number of specific models. For each simulation, a specific model produced between 8,000 and 10,000 utterances, depending on the simulation. The one-stage model had two lexical to phonological and two phonological to subphonemic information flow options. This resulted in four information flow combinations, all but one of which included feedback between at least two levels. Combined with the parameter settings, there were 20,412 specific one-stage models. Of these, 5832 shared the same processing stage and information flow architecture as Dell's (1986) original implementation. The two-stage model has two lexical to phonological and four phonological to subphonemic information flow options. This resulted in eight information flow combinations, five of which included feedback. Combined with the parameter settings, there were 37,908 specific two-stage models.

Such large scale simulations are possible with cluster computing technology. This permits specific models to be run in a massively parallel fashion by distributing the tasks across a network of computers. Using the Edinburgh Compute and Data Facility (ECDF; University of Edinburgh, 2007), we split the simulations into 108 chunks, which normally allowed us to run a study in less than 8 hours. As no interaction is required with the simulations once they have been submitted, it was possible to complete studies overnight.

The models were implemented in Java 1.5. Some studies required statistical comparisons of output between conditions (e.g., for Goldrick and Blumstein's 2006 data, VOTs of intended and unintended productions were compared). As such a large number of specific models were tested, it was advantageous to run these comparisons on the fly. Some such comparisons were carried out within Java, and others by linking Java to R (R Development Core Team, 2008). All in all, the implementation and unit tests comprise 35,000 lines of code.

3.8 Chapter summary

This chapter outlined the representations used in the implementation of one-stage phonological encoding and two-stage phonological encoding and subphonemic processing models. We described how the output was interpreted in simulations of transcribed speech errors and newer instrumental evidence. Manipulations of information flow between levels were described, and the multiple parameter settings used to create specific models and their relationship to settings used in the previous literature were outlined. Finally, the practical details of the implementation were discussed, with particular emphasis on our approach to testing the implementation at multiple parameter settings.

Chapter 4

Effects of parameter manipulations on basic model behaviour

4.1 Introduction

In this chapter, we build a foundation for our later simulations by increasing our understanding of how the basic behaviour of Dell's (1986) original one-stage phonological encoding model is affected by manipulations of the spreading activation parameters. We selected two simple measures by which to analyse the models' behaviour. Firstly, we examined the overall *error rate* of specific models. Secondly, we investigated to what extent errors are contextual, i.e., involve misordered productions of other parts of the utterance, as is the case in anticipations, perseverations and exchanges. In our analyses, we refer to the *non-contextuality* of the errors, so that high readings mean that the network is behaving more randomly for both of our measures.

The data that the present thesis focuses on all relate to errors in speech. However, humans do maintain some level of accuracy in their speech production. Corpus evidence also strongly suggests that when humans do make errors, these errors are frequently contextual (e.g., del Viso et al., 1991; Pérez et al., 2007; Shattuck-Hufnagel & Klatt, 1979; Vousden et al., 2000). A good model of word production must therefore only generate a limited number of errors, and those errors which are generated should exhibit the tendency towards contextuality observed in human productions. A second aim of this chapter was therefore to propose acceptable upper limits on error rate and non-contextuality of errors using corpus and experimental data, and to investigate which parameter settings leave the model able to observe these limits.

These initial investigations also provide an opportunity to explain the graphical approach to examining the effects of parameters which is used throughout this thesis, alongside the regression modelling which adds statistical weight to our graphical observations.

4.2 Simulation methodology

In this section, we outline the configuration of the model used in this simulation, the lexicon of the model, and the task which the model carried out, and we explain how output of the model was interpreted.

4.2.1 Model configuration

In this initial investigation, we focused on evaluating the effect of varying the parameter settings within a model with the same processing stage and information flow settings used by Dell (1986). In other words, a one-stage phonological encoding model with output from the phoneme level was used, with feedback from phonemes to words, and feedback from features to phonemes.

Using only one processing mode and one set of connectivity settings, combined with all specified parameter settings (as outlined in section 3.6) resulted in 5832 specific models.

4.2.2 Model lexicon

The lexicon for this simulation consisted of 48 CVC words selected from the British English Example Pronunciation (BEEP) dictionary (Robinson, n.d.), plus the words *gap* and *cap*. These last two words were included as the same lexicon was used for a later simulation of Goldrick and Blumstein's (2006) data as described in chapter 7, in which we employed *gap* and *cap* as target words. Words for the lexicon were selected to observe the following constraints. Only words with one known pronunciation were allowed, and vulgar words were not permitted. Furthermore, no words where the vowel was a diphthong were included, to simplify vowel representation in the network. Finally, to ensure word level influences on the production of the /g/ and /k/ in *gap* and *cap* in the later simulation of Goldrick and Blumstein's (2006) data, we required there to be at least one more word in the lexicon beginning with /g/, and at least one more beginning with /k/. The resulting lexicon is presented in table 4.1.

Table 4.1: The lexicon of the model for the current simulation.

barb	cull	jam	love	source
bard	feat	jeff	luck	sup
beep	fen	kick	mart	tarn
bun	gap	kit	park	teak
cad	gob	knock	pause	tech
cap	hack	knot	pig	teethe
carl	hag	lad	pod	thug
cease	harsh	lash	root	wring
cell	hock	lid	rot	youth
come	horn	loll	seed	zoom

4.2.3 Task

The model's task for this simulation was to produce pairs of words randomly selected from the lexicon. Each simulation produced 10,000 word pairs. The list of 10,000 word pairs was generated in advance of running the simulations, and the same list was used for all simulations. The single constraint on word pair selection was that the words in a pair could not begin with the same onset, so that contextual onset errors were possible. As the lexicon contained only 50 words, most word pairs occurred in the list more than once.

Word pair production began with a jolt of activation to the first word in the pair, and priming of the second word in the pair. After the specified number of timesteps, the most activated phoneme at each syllable position was selected. Following selection, the activation level of the first word and selected phonemes was set to zero. The activation level of the second word in the pair was then incremented by the jolt activation, and processing continued for the specified number of timesteps. To complete production of the word pair, phoneme selection for the second word then occurred at each syllable position.

We note that one aspect of this procedure may have inadvertently differed from the word pair production procedure used by Dell (1986). In Dell's (1986) theory, selection at any level involves the most activated node being selected and marked for the appropriate location in the frame which is current being filled. The activation level of the selected node is then set to zero, to prevent repeated selection of the node. Selection occurs as many times as is necessary to fill the frame. Processing at the next level involves the representations at the original level which have been placed in the frame sequentially being marked as the *current* representation, following the order specified by the frame. The *current* representation is given a jolt

of activation. This activation flows to connected representations at the following level, where selection proceeds as previously described. Once selection has been completed, the next node in the frame at the original level is marked as the *current* representation and is jolted.

In our implementation, the activation level of the previously *current* node, which is in this case always at the word level, is set to zero when it is unmarked as *current*. This reduces potential interference with subsequent processing caused by activation from this node. However, we suspect that in Dell's (1986) original implementation the activation level of the previously *current* node was not set to zero. This potential difference was noticed at a late stage. It seems that both approaches are reasonable. If there is a mechanism which can mark and possibly unmark nodes as *current* and add jolts of activation to *current* nodes, and if mechanisms for inhibition of activation on selected nodes also exist, it would not seem out of place for the mechanism to also suppress activation from previously current nodes. This is likely to improve accuracy of the network as noted above. Other authors have also observed that inhibition of the word node may follow the production of the phonemes in a word (Schade & Berg, 1992). However, in the following analyses we attempt to highlight points at which this potential unintended difference may have affected our results, to try and understand the influence of this design decision on the model's behaviour.

4.2.4 Onset error classification

Onset productions could be classified as correct productions, contextual errors, or non-contextual errors. A correct production was recorded if the onset produced was the intended onset. A contextual error was recorded if the onset produced was the intended onset of the other word. A non-contextual error was recorded if the onset produced was neither the intended onset, nor the intended onset of the other word.

On the basis of these classifications, the error rate of a given simulation and the proportion of errors which were non-contextual were calculated. The error rate was defined as the percentage of completed onset productions (i.e., productions which were not aborted due to zero selections, see section 3.4.4) which did not result in correct productions. The proportion of non-contextual errors was defined as the percentage of erroneous onset productions which were non-contextual errors. If there were no errors at all for a simulation this proportion could not be calculated.

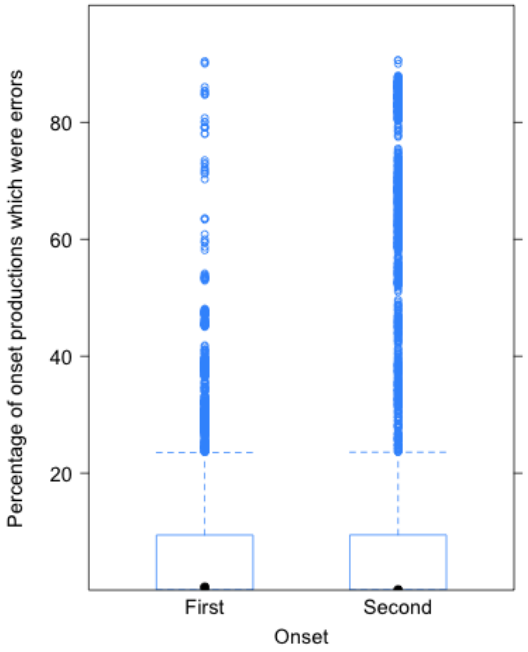


Figure 4.1: Error rate on the first and second onset, across all specific models.

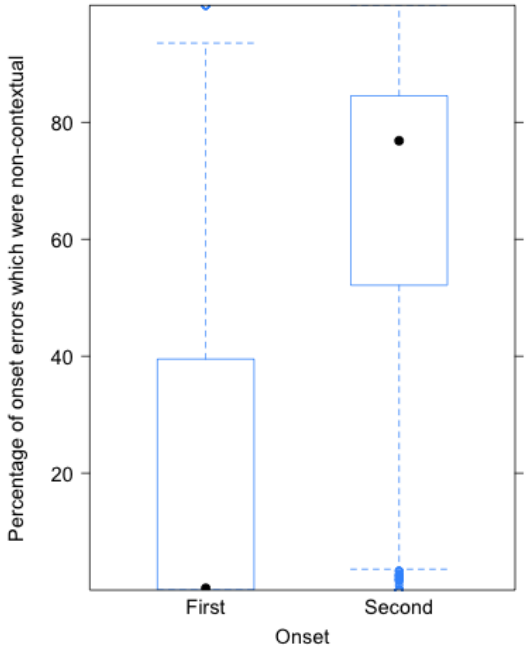


Figure 4.2: The proportion of errors which are non-contextual at the first and second onset. This proportion can only be calculated for specific models which generated at least one error on the specified onset.

We calculated these measures for productions at the first onset only, productions at the second onset only, and finally for productions at both onsets combined.

4.3 Overview of implementation's behaviour on the first and second onset

This section examines and compares the behaviour of our implementation on the onsets of the first and second words of the word pairs produced. We outline the variation of error rate and proportion of errors which were non-contextual for each onset, across all of our specific models.

Figure 4.1 shows that our manipulations of the spreading activation parameters resulted in wide variation of error rates across our 5832 specific models. Error rates on the first onset spanned from 0.00% to 90.47%, and second onset error rates had a similar range, from 0.00% to 90.63%. The median error rates were also alike, with

a median of 0.48% errors on the first onset, and 0.02% errors on the second onset.[1] However, the graph shows that more specific models displayed a tendency towards very high second onset error rates than towards very high first onset error rates. For example, the 90th quantile of the error rate for the first onset was 23.15%, compared to 64.47% for the second onset.

The proportion of non-contextual errors generated can only be calculated for specific models which generate at least one error on the specified onset. Of our 5832 specific models, 3878 generated at least one error on the first onset, and 3187 generated at least one error on the second onset. From figure 4.2, we can see that the proportion of errors which were non-contextual also differed greatly between specific models, with a minimum of 0% and a maximum of 100% of errors being non-contextual for both the first and second onset. However, whilst most specific models generated many more contextual than non-contextual errors on the first onset, with a median of only 0.32% first onset errors being non-contextual, the majority of parameter settings resulted in high proportions of non-contextual errors on the second onset, as reflected by a median second onset non-contextual error proportion of 76.85%.

It is not entirely surprising that more non-contextual errors are generated on the second onset, as contextual first onset errors are directly primed, whereas contextual perseveration errors on the second onset result from the previously produced and suppressed phoneme receiving activation from neighbouring representations in the network, such that activation must be much more dispersed throughout the network in this case, and non-contextual representations are likely to have more activation than they would during first onset production. The tendency of the implementation to produce quite such high proportions of non-contextual errors on the second onset may be cause for concern however. In section 4.5 we return to this point and establish upper limits on the production of non-contextual errors in human productions. Furthermore, somewhat surprisingly, Dell (1986) reports that no non-contextual errors were generated in his original simulations, and we return to this result in section 4.4.3.

[1] We focus on medians rather than means as a measure of central tendency as these error rate and non-contextuality distributions are distinctly non-normal, as is clear from the boxplots.

4.4 The effect of manipulating parameters on the basic behaviour of the implementation

Having established the range of error rates and proportions of non-contextual errors demonstrated by the specific models across the first and second onset, we examined the effect of the parameter manipulations on these measures. This section begins by outlining the statistical and graphical approaches used in our parameter manipulation investigations. The direction and size of each of the effects of the parameters on the error rate and non-contextuality of errors on the first and then second onset is subsequently explored using this methodology, and some reasons for these effects are proposed.

4.4.1 Analysing the effects of manipulating spreading activation parameters

Analysis of the effects of manipulating spreading activation parameters was based on a combination of graphical methods and statistical analyses. These two analysis tools are outlined here, beginning with the statistical component of this analysis.

Regression analysis

In this thesis, the effects of the parameter manipulations on various aspects of the behaviour of the implementation were modelled using regressions. In each case, the model behaviour measure of interest was the dependent variable, and the parameters were the independent variables, apart from adjustments made for collinearity as explained below.

Because of the constraint that *fbkConn* could not exceed *fwdConn*, as explained in section 3.6, *fwdConn* and *fbkConn* were collinear (lower values of *fwdConn* required lower values of *fbkConn*). This correlation makes it difficult to statistically assess the independent contributions of these two parameters, and analyses which attempt to do so may produce misleading results. To avoid this issue, we created a new parameter *connectivity* to represent the joint effect of the two connection strength parameters and replace these two parameters in the logistic regression analysis. As either *fwdConn* or *fbkConn* increase, *connectivity* also increases, and thereby represents the overall increase in activation flow through the network, regardless of the direction of the flow. Specifically, *connectivity* is calculated using the following equation:

Table 4.2: The relationship of *connectivity* values to *fwdConn* and *fbkConn* values

fwdConn	*fbkConn*	*connectivity*
0.05	0.05	0.0025
0.2	0.05	0.01
0.35	0.05	0.0175
0.2	0.2	0.04
0.35	0.2	0.07
0.35	0.35	0.1225

$$connectivity = fwdConn \times fbkConn$$

The *connectivity* values which result from this equation given the *fwdConn* and *fbkConn* values tested in our simulations (see section 3.6) are provided in table 4.2. For models which include feedback, there is an equal number of specific models at each *connectivity* value, because each of these values represents a unique combination of a *fwdConn* value and a *fbkConn* value.

A similar approach was taken to dealing with the *jolt* and *prime* parameters. As *jolt* had to be greater than *prime*, as explained in section 3.6, *jolt* and *prime* were also collinear (lower *jolt* values required lower *prime* values). We again constructed a parameter to represent the combined effect of these two parameters, called *joltPrimeRatio*. The effect of the *prime* parameter is clearly dependent on the *jolt* setting. Values of *prime* higher than the selected *jolt* value are not permitted as they would leave the network unable to reliably produce the target representation rather than the primed representation. Presumably the effect of *prime* values lower than the selected *jolt* value is similarly largely determined by how large a proportion of the *jolt* value they represent. Whilst the *jolt* parameter plays a role in setting the activation scale of the network (Dell, Schwartz, et al., 1997), larger *jolt* settings have also been shown to determine how activated a selected representation is in comparison to unselected competitors (Goldrick, 2006; Rapp & Goldrick, 2000). The size of the *prime* interacts with this latter role of the *jolt* parameter as larger *prime* settings will increase the activation level of competitors. In this instance therefore, the parameter we constructed, *joltPrimeRatio* denoted the relative size of the *jolt* given the *prime*. Specifically, *joltPrimeRatio* was calculated according to the following equation:

$$joltPrimeRatio = \frac{jolt}{prime}$$

Table 4.3: The relationship of *joltPrimeRatio* values to *jolt* and *prime* values

jolt	*prime*	*joltPrimeRatio*
150	100	1.5
200	100	2
100	50	2
150	50	3
200	50	4
50	10	5
100	10	10
150	10	15
200	10	20

The *joltPrimeRatio* values which result from this equation given the *jolt* and *prime* values tested in our simulations (see section 3.6) are provided in table 4.3.

All *joltPrimeRatio* values result from a single combination of one *jolt* and one *prime* value, apart from the *joltPrimeRatio* of 2, which is formed from both a *jolt* value of 200 with a *prime* value of 100, and a *jolt* of 100 with a *prime* value of 50. Whilst we did not choose the *jolt* and *prime* parameters with the creation of the *joltPrimeRatio* parameter in mind, our choice of parameter settings was influenced by the observation that in all previous simulations which have used a *prime* (Dell, 1986; Hartsuiker, 2002), the *jolt* has been twice as big as the *prime*.

All other parameters (*decay*, *steps*, *actiNoiseSD*, *intrinNoiseSD*) were directly added to the regression models as predictors, alongside *connectivity* and *joltPrimeRatio*.

In this chapter, as in many parts of this thesis, the dependent measures were all error based, and therefore binary: either a given error occurred, or it did not. The regression model used for such error based measures was a logistic regression. For logistic regression models, we summarise the Wald's Z calculated from each coefficient and its estimated standard error (Agresti, 2002), in order to explore the comparative size of effects and evaluate the importance of each parameter manipulation for every measure. The contribution of each predictor to a model was estimated using likelihood ratio tests (LRT) and chi-squared tests on the likelihood ratio tests (Agresti, 2002), and the LRT and results of the chi-squared tests are also provided. Example results from regressions of parameter effects on first onset error rate and non-contextuality of first onset errors can be seen in table 4.4.

In this chapter, tests on the likelihood ratio tests showed that all parameters always made highly significant contributions ($p < 0.0001$) to models of both error rate and

non-contextuality. It is perhaps unsurprising that these extremely basic measures of the behaviour of the network are so sensitive to parameter manipulations. The activation spreading calculation dictates that every parameter has an effect on activation distribution throughout the network, and for the error rate regressions, every specific model is measured 10,000 times on the first onset and 10,000 times on the second onset, such that these analyses have great power to detect small effects. Whilst the likelihood ratio test and chi-squared results are given for reference, the examinations of the parameter effects on error rate and non-contextuality presented in this chapter do not refer any further to these significance tests, focusing instead on the direction of the effect and the size of the Wald's Z value. However, the significance of an effect is considered in section 4.6, when we examine which specific models are affected by zero selection problems, as some parameters do not have significant effects on this measure, probably as a result of the extremely low number of zero selections which occur.

Graphing the effects of parameters

All the figures in this thesis which illustrate the effects of spreading activation parameter manipulations on an aspect of model behaviour are presented in a similar manner. Here, we explain this format to aid interpretation of these graphs.

Figure 4.3 is a typical example of a graph used to illustrate the effect of parameters on the behaviour on the network. Each of the graphs in figure 4.3 shows boxplots of the error rate of all the specific models tested at each of the settings of a particular parameter. For instance, the graph at the bottom left of figure 4.3 shows boxplots of the error rate of specific models at each of the decay rates tested. The rest of the figure comprises similar graphs for each of the other parameters. As the values of *fbkConn* tested are dependent on the value of *fwdConn*, boxplots of the error rate of specific models at different *fwdConn* settings are displayed separately for each *fbkConn* setting, and vice versa. The same applies for *jolt* and *prime*.

The layout used for this figure is common to all of the parameter analysis figures in this thesis. The top row of the figure contains two graphs: the first showing the behaviour of specific models at different *fwdConn* settings, displayed separately for each *fbkConn* setting, and the second showing the behaviour of specific models at different *fbkConn* settings, displayed separately for each *fwdConn* setting. The second row contains one graph showing the behaviour of specific models at different *connectivity* values, where the *connectivity* value is calculated from the *fwdConn* and *fbkConn* settings as explained in section 4.4.1. Using these graphs, the results

of the regression analyses demonstrating the effect of manipulating *connectivity* can be related back to the original connection strength variables *fwdConn* and *fbkConn*.

The third row contains two graphs: the first showing the behaviour of specific models at different *jolt* settings, displayed separately for each *prime* setting, and the second showing the behaviour of specific models at different *prime* settings, displayed separately for each *jolt* setting. The fourth row contains one graph showing the behaviour of specific models at different *joltPrimeRatio* values, where the *joltPrimeRatio* value is calculated from the *jolt* and *prime* settings as explained in section 4.4.1. This combination of graphs again helps links to be drawn between the regression analyses on the effect of *joltPrimeRatio* and the influence of the original parameters *jolt* and *prime*.

The fifth and final row contains four graphs. The first graph shows the behaviour of specific models at different *decay* settings; the second shows the behaviour of specific models at different *steps* settings; the third shows the behaviour of specific models at different *actiNoiseSD* settings; and the fourth shows the behaviour of specific models at different *intrinNoiseSD* settings.

Unless stated otherwise in the caption, every graph in each such figure includes every specific model tested in the simulation. Each graph offers a different view of the same data by splitting the specific models up according to the settings of different parameters.

4.4.2 Effects of parameter manipulations on first onset behaviour

To recap, in this simulation, the specific models tested were all parameter setting variations of a one-stage phonological encoding model based on Dell's (1986) implementation, with feedback from phonemes to words, and feedback from features to phonemes. Figure 4.3 depicts the effect of parameter manipulations on the first onset error rate of all the specific models we tested, and figure 4.4 shows how parameter manipulations affect the non-contextuality of the first onset errors generated by the specific models. Table 4.4 shows the results of two logistic regressions, analysing the effect of the parameter manipulations on the first onset error rate, and on the non-contextuality of the errors.

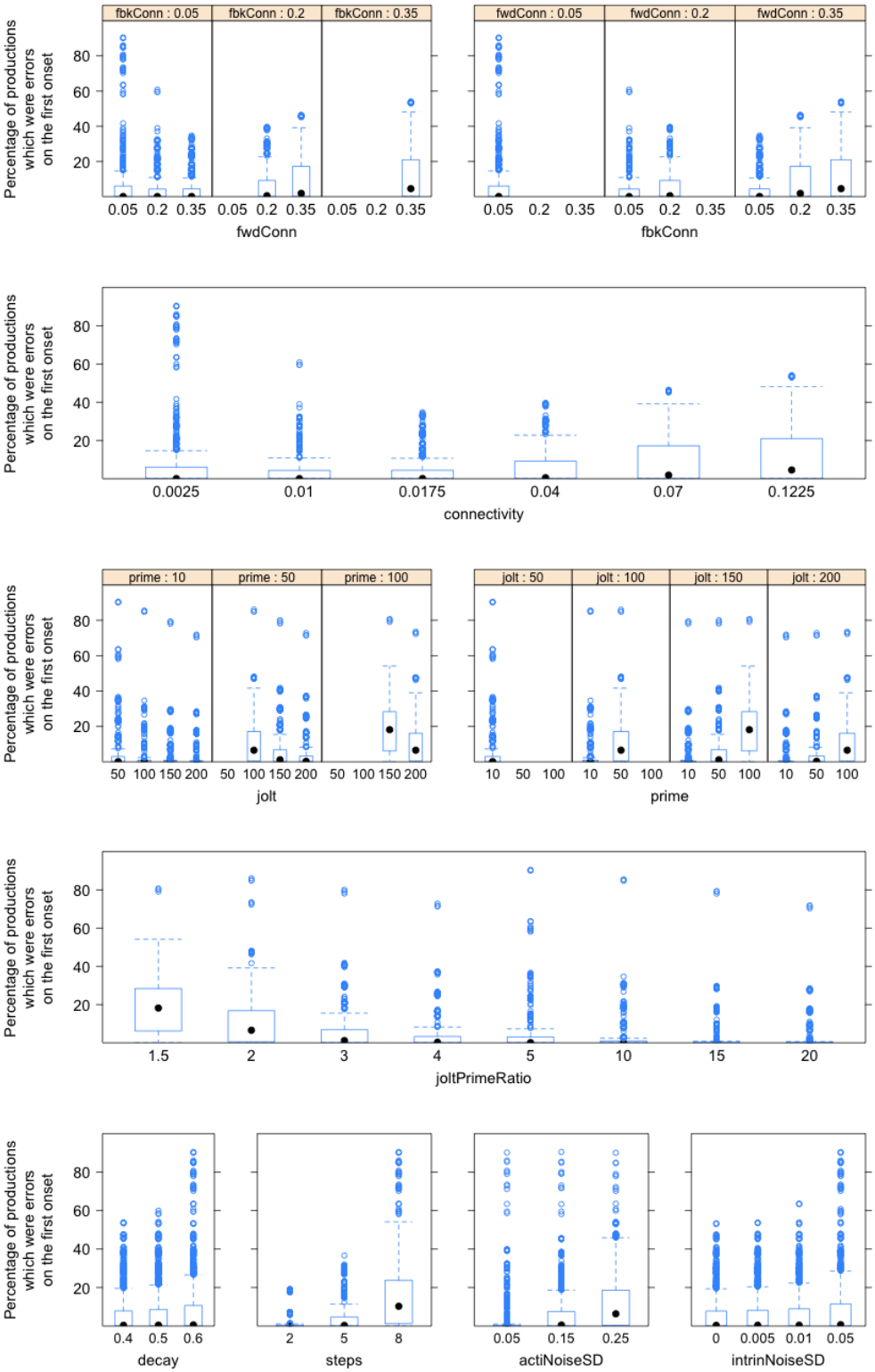


Figure 4.3: The effect of changing parameter settings on first onset error rate, for all specific models.

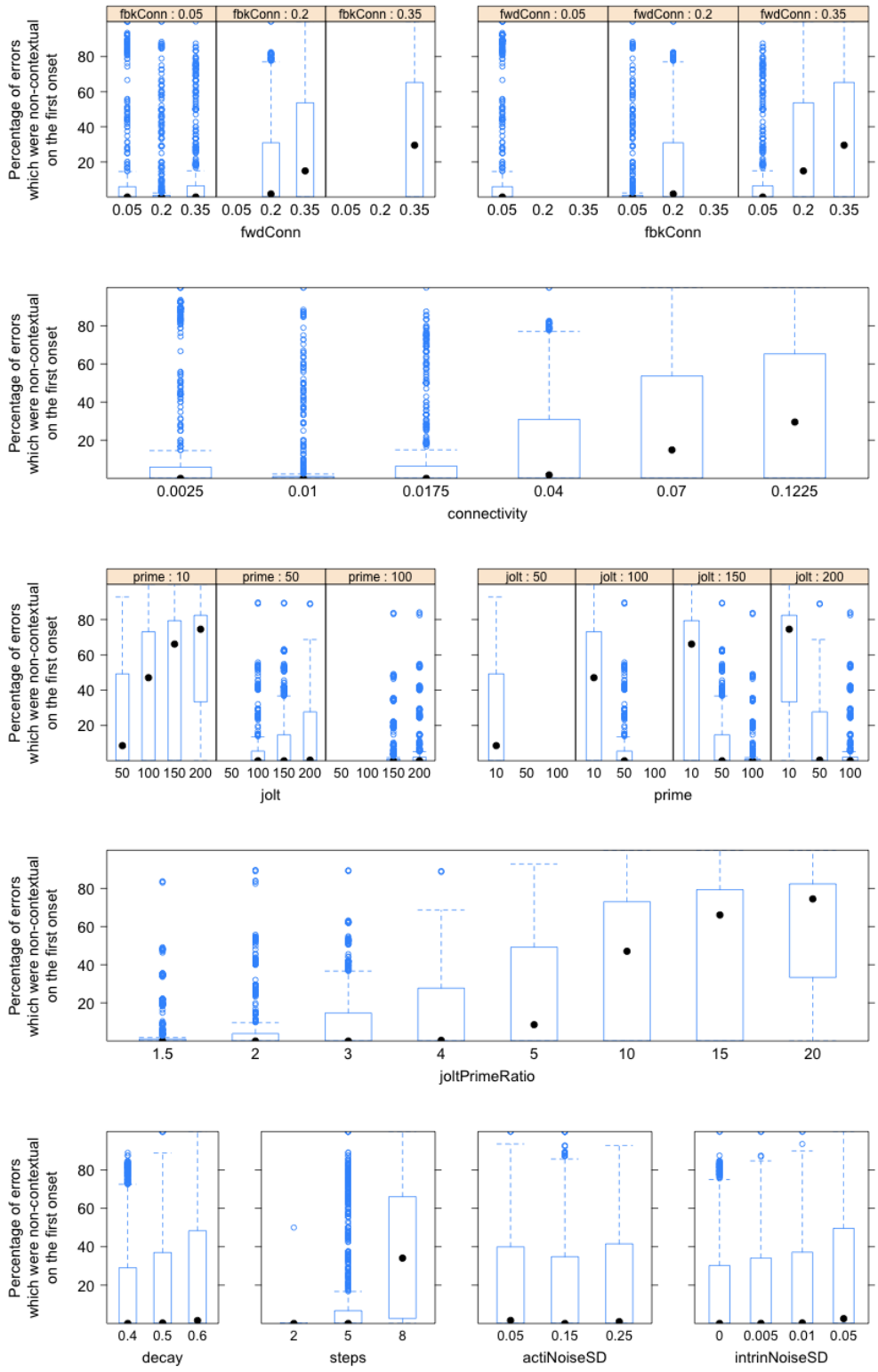


Figure 4.4: The effect of changing parameter settings on the proportion of errors which are non-contextual at the first onset. This proportion can only be calculated for specific models which generated at least one error.

Table 4.4: Results of logistic regression model analyses using parameter values to predict error rate and proportion of errors which were non-contextual on the first onset. The proportion of errors which were non-contextual can only be calculated for specific models which generated at least one error on the specified onset. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	**Error rate**					**Non-contextuality**				
	Dir	Z	LRT	P (χ^2)		Dir	Z	LRT	P (χ^2)	
connectivity	+	821.3	658935	< .001	*	+	321.8	107182	< .001	*
joltPrimeRatio	−	996.9	1555950	< .001	*	+	532.7	450077	< .001	*
decay	+	233.6	54833	< .001	*	+	221.5	50215	< .001	*
steps	+	1519.8	3158908	< .001	*	+	459.1	489315	< .001	*
actiNoiseSD	+	1051.0	1231872	< .001	*	+	133.6	18115	< .001	*
intrinNoiseSD	+	301.6	88513	< .001	*	+	280.2	79399	< .001	*

Key:
Dir = direction

Connectivity: forward connection strength and feedback connection strength

The results of the logistic regressions presented in table 4.4 suggest that as connection strength gets stronger, first onset error rates increase and a higher proportion of non-contextual errors are generated on the first onset. Figures 4.3 and 4.4 confirm that at the higher forward and feedback connection strengths tested, specific models do tend to generate more errors, and these errors are more likely to be non-contextual.

However, closer examination of figure 4.3 shows that some specific models with very low forward and feedback connection strength also generate a very large number of first onset errors, whereas this is true for fewer specific models when forward connection strength is increased. Medians of the distributions of error rate and non-contextuality of errors across specific models are all 0% when *fbkConn* is 0.05 and *fwdConn* is 0.05 or 0.2. However, where *fwdConn* and *fbkConn* are both 0.05, making *connectivity* 0.0025, the upper quartile of the first onset error rate measure is 6.00% errors, whereas when *fwdConn* increases to 0.2, making *connectivity* 0.01, the upper quartile of the error rate is slightly lower at 4.36%. Furthermore, 3.4% of the specific models tested with a *fbkConn* and *fwdConn* setting of 0.05 produce errors on over 50% of their first onsets, whereas this is true for only 0.3% of specific models tested when *fwdConn* is increased to 0.2.

Figure 4.4 further demonstrates that a very low connection strength increases the tendency of specific models to generate non-contextual errors. Specifically, when both *fbkConn* and *fwdConn* are 0.05, the upper quartile of the first onset non-contextuality measure is at 5.80% error non-contextuality, whereas when *fwdConn* is increased to 0.2, the upper quartile of the non-contextuality measure is much lower at 0.95%.

Whilst an increase of *fwdConn* from 0.05 to 0.2 when *fbkConn* is 0.05 results in a decrease of the number of specific models which generate a large number of errors and a high proportion of non-contextual errors, our graphs and further examination of the upper quartiles of these distributions suggest that increasing *fwdConn* can also reduce the accuracy of the network. Increasing *fwdConn* from 0.2 to 0.35 when *fbkConn* is 0.05 leads to a small increase in the upper quartile of the error rate distribution from 4.36% to 4.44% (although the maximum of the error rate distribution decreases from 90.47% to 34.77%). When *fbkConn* is 0.2, increasing *fwdConn* from 0.2 to 0.35 leads to a larger increase in the upper quartile of the error rate distribution, from 9.19% to 17.21% (and here the maximum also increases from 39.64% to 46.50%). A similar pattern emerges when examining the non-contextuality of errors on the first onset. Increasing *fwdConn* from 0.2 to 0.35 when *fbkConn* is 0.05 leads to an increase in the upper quartile of the first onset error non-contextuality distribution from 0.95% to 6.28%. When *fbkConn* is 0.2, increasing *fwdConn* from 0.2 to 0.35 leads to a large increase in the upper quartile of the non-contextuality distribution, from 30.83% to 53.53%.

At the parameter settings tested here, raising the feedback connection strength generally reduces the accuracy of the network. When *fwdConn* is set to either 0.2 or 0.35, an increase to *fbkConn* leads to higher median error rates and higher median proportions of non-contextual errors on the first onset. When *fwdConn* is 0.2, an increase in *fbkConn* from 0.05 to 0.2 results in an increase of median error rate from 0.06% to 0.54%, and an increase of median proportion of non-contextual errors from 0% to 1.75%. When *fwdConn* is 0.35, an increase in *fbkConn* from 0.05 to 0.2 to 0.35 results in an increase of median error rate from 0.10% to 1.87% to 4.54%, and an increase of median proportion of non-contextual errors from 0% to 14.90% to 29.52%.

These results suggest that both low and high connection strengths generally cause higher error rates and higher proportions of non-contextual errors. At medium connection strengths, the tendency for specific models to generate many errors and to generate errors at random is reduced. Specifically, in our data, if forward

and feedback connection strengths are very low, an increase in forward connection strength appears to improve the accuracy of the network, but when either forward or feedback connection strengths are stronger, an increase in forward connection strength begin to reduce the accuracy of the network. At the parameter settings we tested, our graphs suggest that increases in feedback connection strength generally tend to reduce the accuracy of the network.

Problems at low connection strengths fit in with the results presented by Dell, Schwartz, et al. (1997), who showed that lower connection strengths lead to higher error rates and more random errors. As Dell, Schwartz, et al. (1997) explain, connection strengths which are too low prevent activation from being able to pass through the network effectively. More errors occur because target representations are not activated sufficiently, and errors become more random because activation patterns at different levels are less in line with one another. Our graphs suggest that increases in forward connection strength in particular can result in an improvement in accuracy, when forward and feedback connection strengths are very low. It has also previously been suggested that when connection strengths are very low, increases in feedback connection strength may reduce error rate by reinforcing the activation conveyed by the weakened forward connections (Ruml et al., 2000). Whilst the current results do not suggest that the accuracy of the network can be improved by increasing feedback strength, the connection strengths tested by Ruml et al. (2000) were extremely low ($fwdConn < 0.1$ and $fbkConn < 0.001$) and so our parameter explorations may not cover the parameter space in which feedback has a crucial reinforcement role.

Some problems with high connection strengths have been demonstrated in two previous studies. In an investigation of an abstract spreading activation network carried out by Shrager et al. (1987), it was shown that the connection strength of a network must be lower than the decay rate, to avoid the activation levels in the network growing without bound due to feedback connections.However, as the highest connection strength tested in the current simulation is 0.35, and the lowest decay rate 0.4, all of our specific models meet this condition. Goldrick (2006; Rapp & Goldrick, 2000) has demonstrated that in models of aphasic patients, inappropriate error patterns are generated when feedback to or from noisy layers is too strong. These results largely rely on feedback connection strengths being much stronger than the forward connection strengths however, whereas feedback connection strengths are never stronger than forward connection strengths in our specific models.

Feedback connections by their nature channel activation to parts of the network other than the target representations. Representations with connections to many other representations in the network are particularly likely to become highly activated (e.g., Hartsuiker, 2002), regardless of whether their production is intended. We posit that inappropriate activation of representations in the network is therefore caused by the presence of feedback connections in particular, reducing the influence of the jolt and prime activation on target and competing representation. This hypothesis fits in with the observation that for the parameter settings we examined, increasing feedback connection strength increases the error rate and non-contextuality of errors. However, at higher forward and feedback connection strengths, increasing forward connection strength also increases the error rate and non-contextuality of errors. We suggest that in these cases, the effect of feedback is being amplified by stronger forward connections, as these connections reinforce the forward flowing part of feedback loops. The claim that too strong an influence of feedback causes problems in processing is in line with Goldrick's (2006) argument for limited interactivity in word production. The results presented here add to the simulations which he reports, as in our investigation, feedback connection strength is never higher than forward connection strength, and in addition, no representations are subjected to activation noise as strong as that used by Goldrick (2006) to simulate aphasic damage ($actiNoiseSD = 0.7$).

Jolt and prime

The regression models presented in table 4.4 indicate that as the ratio of jolt to prime increases, fewer errors occur, but those which do occur are more likely to be non-contextual. This is confirmed by figure 4.3, in which the effect of *joltPrimeRatio* on the first onset error rate of specific models is shown, and figure 4.4, which depicts the effect of *joltPrimeRatio* on the proportion of non-contextual errors at the first onset.

This result follows from a simple understanding of how the network operates. As the jolt to prime ratio decreases, or in other words, the prime increases proportional to the jolt, the network is more likely to produce the competing phoneme in error, as it will be proportionally more activated in comparison to the target phoneme. At lower jolt to prime ratios, specific models therefore exhibit a higher error rate. This behaviour clearly relates to Goldrick's (2006) observation that jolt size affects how strongly the target representation is selected, and also Dell's (1986) claim that increasing the prime will increase the chance of anticipatory errors.

The jolt also sets the overall activation scale of the network (Dell, Schwartz, et al., 1997). A prime which is proportionally higher relative to the jolt will therefore also mean that the competing phoneme is more active in comparison to other representations which could potentially be produced in error. As a result, specific models with lower jolt to prime ratios generate a higher proportion of contextual errors, or in other words, a lower proportion of non-contextual errors.

Decay

The logistic regressions presented in table 4.4 show that as the decay rate increases, more first onset errors occur, and these errors are more likely to be non-contextual. This is reflected in the boxplot of first onset error rate by *decay* in figure 4.3, and the boxplot of the proportion of non-contextual errors on the first onset in figure 4.4.

Again, these results are intuitive. As Dell, Schwartz, et al. (1997) have emphasised, a higher decay rate impairs the ability of the network to retain activation. This affects both the activation from the target word, therefore increasing the error rate, and the activation from the competitor word, therefore making it more likely that errors are non-contextual.

This appears to be a comparatively weak effect at the settings which we have tested, however. The Wald's Z value for the effect of the *decay* parameter on the first onset error rate is 233.6, the lowest of all parameters, compared to a mean Z value of 820.7. The Z value for the non-contextuality measure is also relatively low at 221.5, compared to a mean of 324.8.

Steps

The statistical analyses in table 4.4 show that as the number of timesteps before selection increases, more first onset errors occur, and these errors are more likely to be non-contextual. Again, this can also be seen in the boxplot of first onset error rate by *steps* in figure 4.3, and the boxplot of the proportion of non-contextual errors on the first onset in figure 4.4. As the Z values in table 4.4 show, the *steps* parameter is by far the strongest predictor of first onset error rate, and nearly the strongest predictor of first onset error non-contextuality too.

In contrast to these results, Dell's (1986) original investigation into the role of the steps parameter claimed that allowing more timesteps to pass before selection will

increase the accuracy of the network's productions, as outlined in section 2.4.1. In the simulations presented by Dell (1986), some errors at low timestep settings were caused by the inability of the jolt activation administered to the target word to reach some target phonemes, as there were more layers between the jolted word and the phoneme than there were timesteps. In these cases, production at selection was not just slightly impaired but entirely random, such that 60% of phoneme productions did not result in the intended phoneme. This effect was very abrupt, such that with one more timestep allowed, no errors occurred at all on the production of the first word. In the current network, no errors are caused in this manner, because the phoneme layer is directly connected to the word layer, as in the majority of other models based on Dell (1986) (Dell, 1986, SLIP task model; Dell, 1990; Dell et al., 2004; Dell & O'Seaghdha, 1991, 1992; Dell, Schwartz, et al., 1997; Foygel & Dell, 2000; Goldrick, 2006; Goldrick & Rapp, 2002; Martin et al., 1994; Oppenheim & Dell, 2008; Rapp & Goldrick, 2000; Ruml et al., 2000, 2005; Schwartz et al., 2006) so one timestep is sufficient for activation to be conveyed.

Some other errors in Dell's (1986) investigations were caused by activation perseverating in the network from previous productions of words. Perseverating activation has a greater effect and causes more errors if there are fewer timesteps before selection, as there is less time for the activation to decay. However, on productions of the first word like the productions we consider in this section, there is no previous production for activation to perseverate from. In models which focus on single word production, errors are never caused by perseverating activation (Dell, 1990; Dell & Gordon, 2003; Dell et al., 2004; Dell & O'Seaghdha, 1991, 1992; Dell, Schwartz, et al., 1997; Foygel & Dell, 2000; Goldrick, 2006; Goldrick & Rapp, 2002; Martin et al., 1994; Rapp & Goldrick, 2000; Ruml & Caramazza, 2000; Ruml et al., 2000; Ruml et al., 2005; Schwartz et al., 2006).

It is not clear from previous investigations however that allowing more timesteps to pass before selection occurs will reduce the number of errors in models where errors are primarily caused by noise. The two key simulations which Dell (1986) used to demonstrate that more timesteps led to fewer errors did not include any activation-based or intrinsic noise. In the one model presented by Dell (1986) which did involve activation-based noise, a very low number of timesteps did increase the number of errors, but the breakdown of error types provided makes it clear that these extra errors were nearly all perseverations, and hence occurred on non-initial words.

Most models which have extended Dell's (1986) theory do include at least activation-based noise, if not also intrinsic noise (Dell & Gordon, 2003; Dell et al., 2004; Dell, Schwartz, et al., 1997; Hartsuiker, 2002; Foygel & Dell, 2000; Goldrick & Rapp, 2002; Martin et al., 1994; Oppenheim & Dell, 2008; Rapp & Goldrick, 2000; Ruml & Caramazza, 2000; Ruml et al., 2000, 2005; Schwartz et al., 2006), and our implementation follows this approach. As noted in 3.6, simulating errors by using noise means that words which the network is capable of producing correctly are sometimes produced erroneously. This contrasts sharply to error generation in networks without noise, where errors are caused by perseverative activation, or because activation cannot travel from the layer at which the jolt is applied to the layer at which selection occurs in the number of timesteps allowed. In these networks, errors which occur on an attempt to produce a specified word in a specified phrase will always occur, such that the network is never able to produce the phrase correctly.

Goldrick and Rapp (2002) have suggested that in a model with aphasic damage simulated by high amounts of activation-based noise ($0.2 \leq \mathit{actiNoiseSD} \leq 0.35$), postponing selection for a greater number of timesteps will increase the number of errors rather than decrease them, as more activation will be able to cascade from the damaged level, and more activation will be able to feedback. In the simulations they report however, they manipulate noise, jolt and steps simultaneously, so it is not possible to see what increase in errors is due to the increase in timesteps before selection.

Here, in models with generally lower amounts of activation-based noise ($0.05 \leq \mathit{actiNoiseSD} \leq 0.25$) than that used in Goldrick and Rapp's (2002) aphasic models, we manipulate the number of timesteps separately and demonstrate that selecting representations after a higher number of timesteps has passed tends to cause specific models to generate more errors and higher proportions of non-contextual errors. Allowing the network more timesteps before selection therefore reduces accuracy, rather than improving it in the way that Dell (1986) suggested. We suggest that using a higher number of timesteps allows more time for the original activation patterns to decay, for feedback connections to activate unrelated representations, particularly those with many connections to other representations, and for noise to gain a bigger influence on the activation patterns, Together, these factors lead to the original activation patterns becoming increasingly obscured. At higher timestep settings, errors are therefore more likely because the influence of the jolt activation is reduced. Errors are also more likely to be non-contextual, because the influence

of the prime activation is similarly reduced. On the assumption that the network makes fewer errors when more timesteps are allowed to pass before selection, Dell (1986) had characterised the number of steps as the inverse of speech rate, where humans are more prone to error when forced to speak fast (MacKay, 1971). We argue that our results demonstrate that in models in which the main cause of errors is noise, the number of timesteps which pass before selection at the first word is perhaps better characterised as the amount of time for which the network has to remember what it is supposed to be saying.

Activation-based noise

The regressions in table 4.4 suggest that as *actiNoiseSD* increases, more first onset errors occur, and these errors are more likely to be non-contextual. Figure 4.3 also shows first onset error rate increasing as *actiNoiseSD* increases, but figure 4.4 shows little effect of *actiNoiseSD* on the proportion of non-contextual errors at the first onset. Another look at table 4.4 reveals that whilst *actiNoiseSD* does have a highly significant effect on the proportion of non-contextual errors, it is the weakest effect of all the parameters for this measure, with a Wald's Z value of 133.6, compared to a mean Wald's Z of 324.8 for parameter effects on this measure. It has one of the strongest effects on first onset error rate however, with a Wald's Z value of 1051.0, compared to a mean of 820.7.

It follows that error rates are higher when larger amounts of activation-based noise are used, as this noise reduces the clarity of the activation patterns in the network. However, the target representations and primed representations are particularly affected, because the amount of noise applied to a representation is proportional to the activation of that node, and these will be by far the most activated nodes to begin with. The strong effect of activation-based noise on the primed representation is reflected in the weak effect of activation-based noise on the non-contextual measure. Our results suggest that blurring the activation patterns in the network by increasing the activation-based noise slightly increases the likelihood of selection of each of the non-contextual phonemes, of which there are many, but particularly increases the likelihood of selection of the primed phoneme, of which there is only one. This slight increase in frequency of production of each of the many non-contextual phonemes combined with a bigger increase in frequency of production of the one primed phoneme balances out to result in only a small increase in the proportion of non-contextual errors.

Intrinsic noise

Finally, we examine the effect of intrinsic noise. The regression results reported in table 4.4 show that as intrinsic noise increases, specific models tend to generate more first onset errors, and higher proportions of non-contextual errors. These effects are reflected in the first onset error rate boxplot in figure 4.3, and the boxplot of proportions of non-contextual first onset errors in figure 4.4.

These results simply reflect the fact that an increase in intrinsic noise reduces the influence of both the jolt activation and the prime activation.

However, at the parameter settings we tested, intrinsic noise has one of the weakest effects on both first onset error rate and non-contextuality of first onset errors, as confirmed by the Z values in table 4.4.

4.4.3 Effects of parameter manipulations on second onset behaviour

Having considered the effect of the parameter manipulations on error rate and non-contextuality of errors at the first onset, we also examined the effect of the parameters on behaviour at the second onset, to illuminate similarities and differences between the direction and sizes of the effects on the two onsets. Figure 4.5 depicts the effect of parameter manipulations on the second onset error rate of all the specific models we tested, and figure 4.6 shows how parameter manipulations affect the non-contextuality of the second onset errors generated by the specific models. Table 4.5 shows the results of two logistic regressions, analysing the effect of the parameter manipulations on the second onset error rate, and on the non-contextuality of the errors.

Connectivity: forward connection strength and feedback connection strength

Both the results of the regression analyses shown in table 4.5 and the graphs of the effects of manipulating *connectivity* on second onset error rate (figure 4.5) and non-contextuality of second onset errors (figure 4.6) show a similar pattern to that seen for the first onset. Again, increasing connectivity increases error rate and non-contextuality of errors. However, as for the first onset measures, figure 4.5 shows evidence of a group of simulations with low *fwdConn* and *fbkConn* settings which exhibit very high error rates, and figure 4.6 similarly shows that a group of simulations with low connection strengths generate very high proportions of non-contextual errors. The effect of connection strength on error rate is clearly

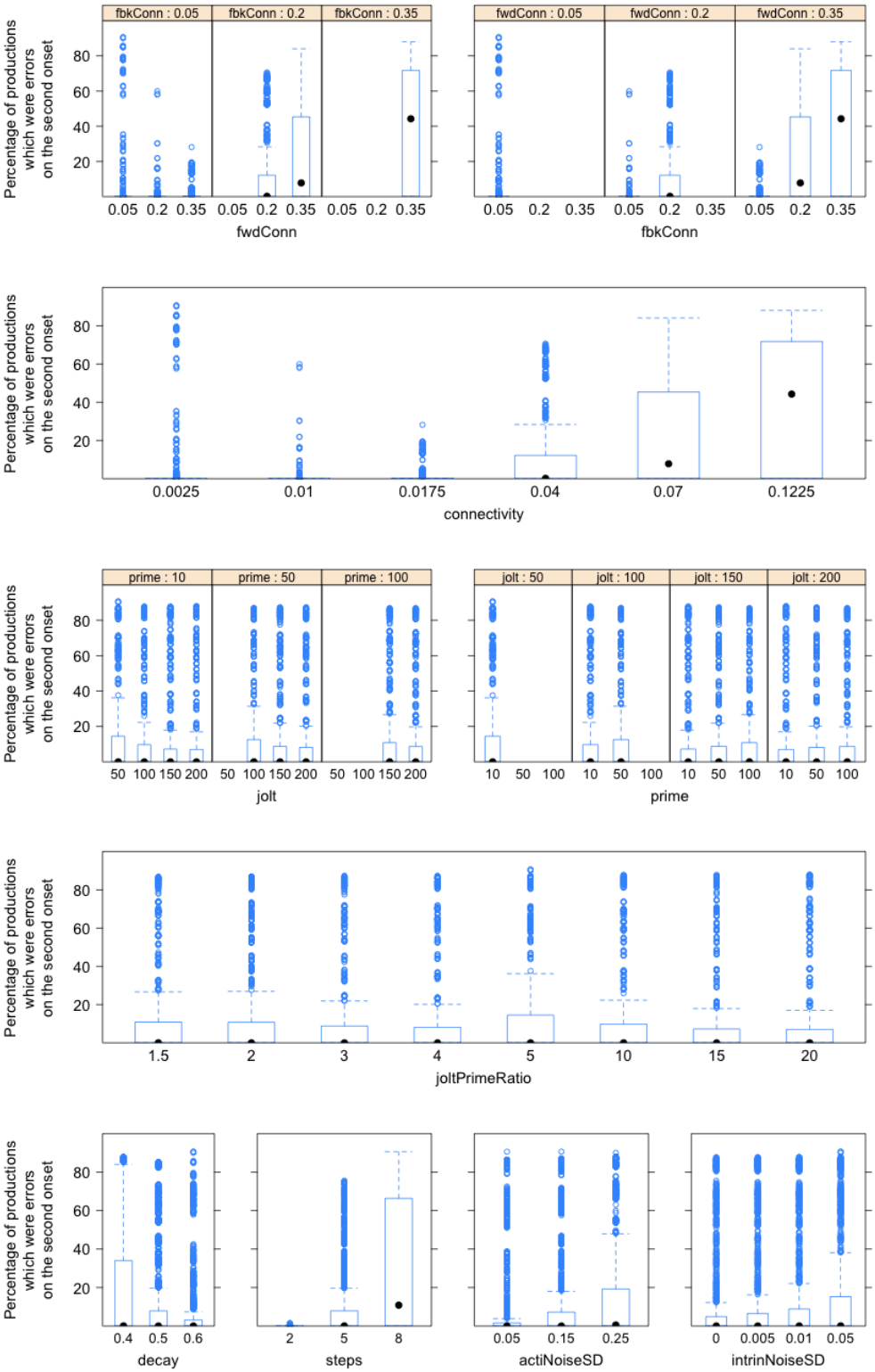


Figure 4.5: The effect of changing parameter settings on second onset error rate, for all specific models.

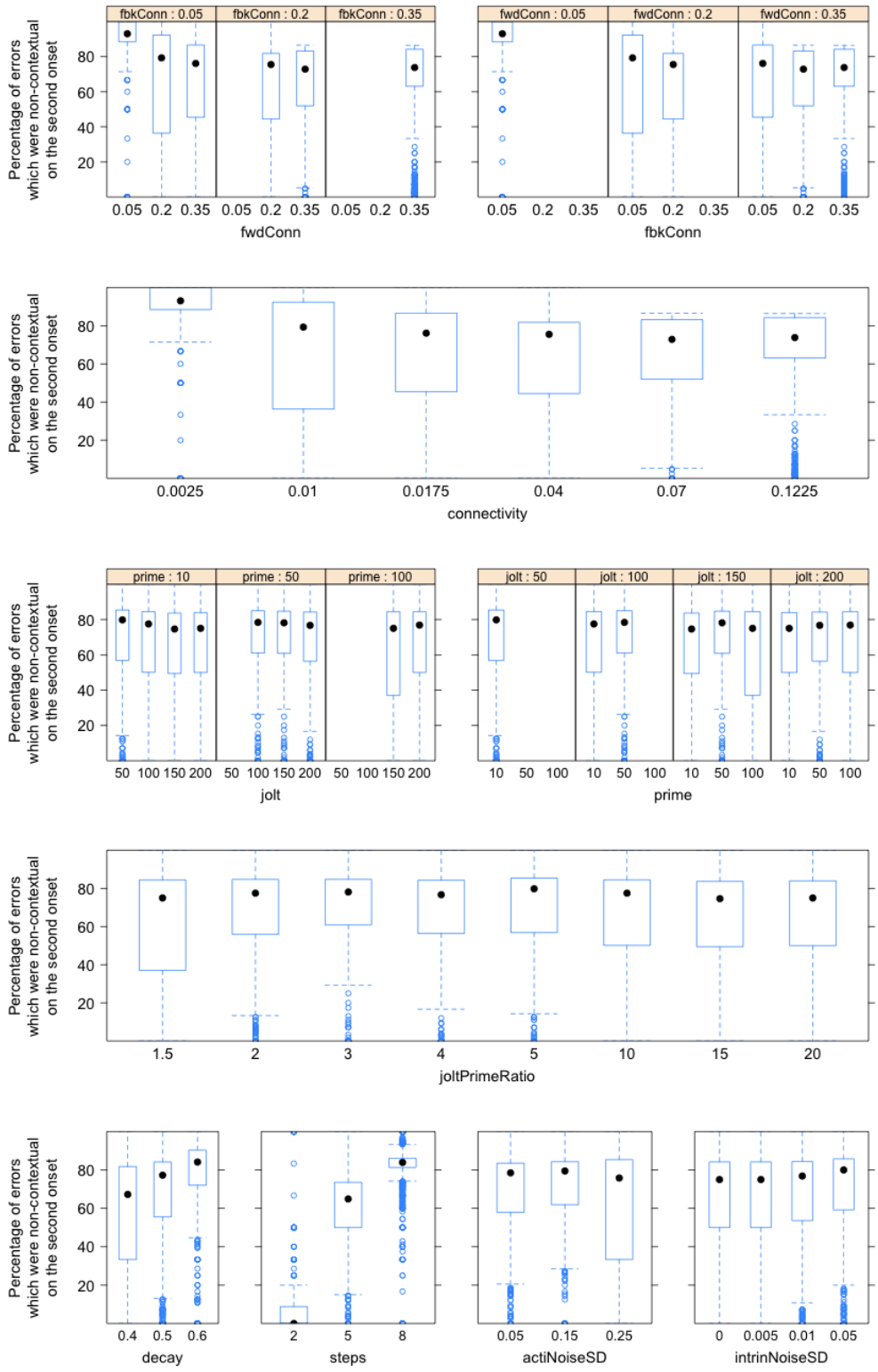


Figure 4.6: The effect of changing parameter settings on the proportion of errors which are non-contextual at the second onset. This proportion can only be calculated for specific models which generated at least one error.

Table 4.5: Results of logistic regression model analyses using parameter values to predict error rate and proportion of errors which were non-contextual on the second onset. The proportion of errors which were non-contextual can only be calculated for specific models which generated at least one error on the specified onset. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	**Error rate**					**Non-contextuality**				
	Dir	Z	LRT	P (χ^2)		Dir	Z	LRT	P (χ^2)	
connectivity	+	2748.3	12812163	< .001	*	+	63.2	3979	< .001	*
joltPrimeRatio	−	41.5	1723	< .001	*	−	87.8	7613	< .001	*
decay	−	1311.3	1962516	< .001	*	+	95.0	9116	< .001	*
steps	+	2411.1	12141160	< .001	*	+	477.1	224116	< .001	*
actiNoiseSD	+	593.4	360917	< .001	*	+	104.8	10976	< .001	*
intrinNoiseSD	+	300.2	89147	< .001	*	+	48.3	2351	< .001	*

Key:
Dir = direction

stronger on the second than the first onset however, as the Wald's Z for the effect of *connectivity* on second onset error rate is 2748.3, the strongest of all parameter effects for this measure, whereas the Wald's Z for the effect of *connectivity* on first onset error rate was 821.3, where three other parameters had higher Wald's Zs. The effect of connectivity on non-contextuality of errors on the second onset was weaker than for the first onset, with a Z value of 63.2 for the second onset, making it the second to least important parameter, compared to 321.8 on the first onset, where it was the third most important parameter for this measure.

We assume that connection strength manipulations affect the processing of the network on the second onset in the same basic manner as was proposed for the first onset. Connection strengths which are too low mean that the second onset jolt activation cannot pass through the network effectively, leading to increased error rates. Low connection strengths also make it unlikely that the first onset will be produced in error, causing a contextual error. To recap, a contextual error on the second onset is either part of a perseveration or part of an exchange. Dell (1986) explains that a perseveration is caused by the selected and suppressed first onset being reactivated during production of the second word. During production of the first word, activation feeds back from the first onset to all words in which the first onset appears in the onset position. For example, in the phrase *big fun*, production of the word *big* will cause activation to flow from the onset /b/ up to the words *bill* and *bat*. Following successful production of the onset /b/, the

onset phoneme will be suppressed. However, during production of the second word, activation can return from *bill* and *bat* and reactivate /b/, on occasion leading it to be reselected as the second onset too. Very low connection strengths would make perseverations unlikely, as the connections needed to activate /b/ in the first place. convey activation to the neighbour nodes *bill* and *bat*, and then pass activation back to /b/ during second word production, would all be impaired. On the other hand, an error on the second onset which is part of an exchange is triggered by an error on the first onset. In the phrase *big fun*, if the /f/ is anticipated and produced in the first onset position, then the /f/ will be suppressed, but the /b/ will not. Remaining activation on the /b/ may cause it to be selected for production in the second onset position instead. However, if connection strengths are too low, /b/ is unlikely to be substantially activated in the first place, making this scenario unlikely.

Connection strengths which are too high lead representations in some specific models to become inappropriately highly activated, particularly representations which have a high number of connections to other representations, regardless of whether they were intended for production. This reduces the influence of the jolt activation, and thereby increases the error rate. This error inducing effect is particularly strong on production of the second word, as activation will have already built up due to high connection strengths during production of the first word. However, the effect of high connection strength on non-contextuality proportions on the second onset is more complicated. Whilst high connection strengths lead unrelated representations to become highly activated, causing generation of non-contextual errors, high connection strengths also support perseverative contextual errors, by firstly increasing the activation conveyed to words such as *bill* and *bat* in which the first onset /b/ participates during first onset production, and then increasing the flow of activation back from these words to the first onset /b/ during second onset production. As a result, the increase in the proportion of non-contextual errors is reduced.

Jolt and prime

The regression analysis summarised in table 4.5 suggests that error rate on the second onset decreases with increasing *joltPrimeRatio*, as is the case for the first onset. However, the strong effect seen on the first onset error rate (Wald's $Z = 996.9$) is not repeated here, and the effect of *joltPrimeRatio* is is the weakest of all effects on the second onset error rate, with a Wald's Z of 41.5, compared to a mean of 1234.3. Figure 4.5 suggests that the independent value of the *jolt* parameter may be important. Whilst the median of the error rate distribution is highest at

joltPrimeRatio = 1.5 (error rate median = 0.05%), the upper quartile of the error rate distribution is clearly highest at *joltPrimeRatio* = 5 (upper quartile = 14.47%), as is the maximum error rate (90.63%). This points to a tendency of some specific models to display higher error rates when *jolt* is at its lowest setting of 50 (which combined with a *prime* value of 10 leads to a *joltPrimeRatio* of 5). Figure 4.5 also shows that the value of *prime* does modulate this behaviour however. As is just about visible in the graph, at each *jolt* setting, the median second onset error rate increases as *prime* increases. It is perhaps surprising that error rate increases rather than decreases as *prime* increases, as we discuss below.

On the first onset, non-contextuality of errors increased as *joltPrimeRatio* increased. The effect of *joltPrimeRatio* was the strongest of all the parameters on this measure, with a Wald's Z of 532.7. On the second onset in contrast, the regression analysis suggests that non-contextuality of errors decreases as *joltPrimeRatio* increases. This parameter is not the strongest predictor of non-contextuality however; whereas the Z value for the strongest predictor *steps* is 477.1, Wald's Z for *joltPrimeRatio* is only 87.8. A closer look at figure 4.6 suggests that there may be a tendency for higher proportions of non-contextual errors at mid-range *joltPrimeRatio* values, rather than at the extremes. The median proportion of non-contextual errors both when *joltPrimeRatio* is 1.5 and when *joltPrimeRatio* is 20 comes to 75.0%, and is slightly lower at 74.6% when *joltPrimeRatio* is 15. Between *joltPrimeRatio* values of 2 to 10 however, the median proportion of non-contextual errors ranges from 76.7% to 79.8%.

It is unsurprising that there is a difference between the effect of jolt to prime ratio on the first and the second onset. On the first onset, priming activation was applied to the upcoming second onset. On the second onset in a phrase of two words, there is no upcoming onset for prime to be applied to. However, effects of processing on the first onset, including the priming of the second onset, may still be felt during processing of the second onset due to perseveratory activation in the network.

We suggest that error rate decreases as the jolt to prime ratio increases due to effects of both the jolt and the prime. As the jolt activation increases, target phonemes are more strongly activated. Increased activation of target phonemes in comparison to other phonemes leads the network to generate less errors. However, the increase in activation compared to the activation of other phonemes in the network is less here than it was on the first onset, as the scale of activation in the network was set by the same jolt size on the production of the previous word. Higher jolts therefore

lead to there being more activation already present in the network for the jolt at the second onset to compete with.

There are two reasons why the error rate may increase as the prime activation increases. Firstly, higher primes are more likely to cause errors on the first onset (as confirmed in section 4.4.2). If an error occurs on the first onset, the intended first onset will not be suppressed, and therefore will be more likely to be erroneously selected at the second onset, creating an exchange. Secondly, higher primes will lead to higher levels of activation in the network overall. Feedback will lead to this activation spreading from the primed onset through the network, and may cause phonemes other than the intended second onset to have sufficient activation to be selected during production of the second word. The fact that both of these explanations are mostly oriented around processing at the previous word may help explain why this effect is rather weak.

As previously noted however, it is to some extent counter-intuitive that increasing the priming of the second onset during production of the first onset reduces accuracy of production of the second onset. Dell, Burger, and Svec (1997) make the seemingly logical argument that production systems should prime upcoming representations with activation to aid their timely retrieval. In contrast, the current results suggest that increasing the priming of an upcoming representation can under some circumstances hinder production of that representation.

Finally, as demonstrated by the higher number of specific models generating high error rates at the lowest jolt setting, on the second onset the role of the jolt size increases in importance in comparison to the role of the jolt to prime ratio. We suggest this is because of the reduced influence of the prime activation due to no upcoming phoneme being primed.

Turning to the non-contextuality of errors on the second onset, we posit that both the overall decrease of non-contextuality of errors as the jolt to prime ratio goes up and the evidence that the highest non-contextuality proportions may be generated at mid-range jolt to prime ratios can be explained by reference to the two different mechanisms proposed by Dell (1986) for generation of contextual errors on the second onset. Consider production of the phrase *big fun*. We suggest that a perseveration error is more likely to occur when the jolt to prime ratio is high. Firstly, a high jolt to prime ratio makes it more likely that the first onset will be produced correctly. At second onset production, the second onset will be less active relative to the rest of the network (in particular, the neighbours of *big*, such as *bill* and *bat*)

than it would have been with a high prime following correct first onset production. Comparatively more activation will therefore pass back to the first onset /b/ during production of the second word, making perseverations more likely.

In contrast, we suggest that a low jolt to prime ratio makes exchange errors more likely. When the prime is high relative to the jolt, anticipations on the first onset are more likely, which leads to the intended first onset not being reset. Once an anticipation has occurred, and the intended second onset has been reset, a low jolt to prime ratio also means that less jolt activation is provided to the intended second onset in proportion to the activation already in the network. Altogether, this means that the probability that the activation remaining on the intended first onset is more than the activation on the intended second onset is therefore increased, leading to more completed exchange errors.

This result fits in with Dell's (1986) claim that a high prime increases the number of exchange errors generated. However, we add to this by noting that this may not only be because of the increased chance of the anticipatory portion of the error. Both through the triggering influence of the first error, and comparatively reduced support provided to the second onset, the jolt to prime ratio also affects exchange error generation at the second onset. We also suggest that a low prime may in fact not only decrease the number of exchanges generated, but increase the number of perseverations too. We will further examine these claims about the effect of jolt to prime ratio on perseverations and exchanges in the following chapter.

To summarise, a low jolt to prime ratio benefits production of contextual second onset errors as part of exchanges, whereas a high jolt to prime ratio causes more perseverations to be generated. A lower proportion of errors are contextual at mid-range jolt to prime ratios, as perseveration generation is reduced in comparison to high jolt to prime ratios, and exchange generation is reduced in comparison to low jolt to prime ratios. The slight overall trend for an increase in the proportion of contextual errors or decrease in the proportion of non-contextual errors as jolt to prime ratio increases can be explained by reference to Dell (1986) results which demonstrate that the model tends to generate more perseverations than exchanges, a result examined in further detail in the next chapter.

Decay

The logistic regression analyses in table 4.5 show that, unlike on the first onset, higher *decay* settings lead to lower error rates. This is also visible in figure 4.5.

Whereas the *decay* was the weakest predictor of first onset error rate, with a Wald's Z of 233.6 (notably less than the mean Z value for parameter effects on that measure, 820.7), the effect of *decay* on second onset error rate is comparatively quite strong, with a Z value of 1311.3 (greater than the mean Z value for parameter effects on second onset error rate, which was 1234.3). Higher *decay* settings also lead to higher proportions of non-contextual errors, as they did on the first onset (see also figure 4.6). As on the first onset, this is a slightly weaker than average effect, with a Z value of 95.0, compared to the mean Z value 146.0.

The higher error rate at lower decay rates is in line with Dell's (1986) observations that lower decay rates cause more perseverations and exchanges to be generated. Both perseverations and exchanges are caused by activation from the production of the first onset remaining in the network. Perseverations occur because words which the first onset participates in (e.g., *bill* and *bat* for the first onset /b/) remain activated and reactivate the first onset during second word production. Exchanges occur because the activation on the first onset, which was not suppressed because the first onset was not selected, remains strong enough for the first onset to be selected in error at second onset selection.

The extra activation in the network at lower decay rates reduces the effect of the jolt activation for the second onset, and leads to more errors. However, whilst overall error rate declines at higher decay rates where this extra activation is weakened, errors have a greater tendency to be non-contextual as there is a lesser influence from production of the previous word.

Steps

The logistic regression analyses summarised in table 4.5 show that, as on the first onset, higher values of *steps* are associated with higher error rates. In line with our first onset results, a higher proportion of the errors at higher values of *steps* are non-contextual. These results are also visible in figures 4.5 and 4.6.

The effect of *steps* is the second strongest effect on second onset error rate, with a Wald's Z value of 2411.1, compared to a mean Z value of 1234.3. On the first onset, *steps* was the strongest predictor, but here it is just beaten by *connectivity* which has a Z value of 2748.3. The importance of the effect is however extremely evident from the fact that no models with a *steps* setting of 2 generate more than 1.48% errors, whereas 25% of models with a *steps* setting of 8 generate over 66.35% errors.

However, whereas *steps* was the second strongest predictor of first onset error non-contextuality, it is by far the strongest predictor of second onset error non-contextuality, with a Z value of 477.1, nearly five times as strong as the next strongest predictor, which is *actiNoiseSD* with a Z value of 104.8.

As observed in section 4.4.2, the result that the error rate increases as more timesteps are allowed before selection is contrary to what Dell (1986) reported. We argued before that the error rate increases with the number of timesteps because a longer number of timesteps causes the jolt activation to decay, leaving noise to play a bigger role in determining activation levels, and creating an opportunity for feedback connections to activate unrelated representations, particularly those with many connections to other representations. Dell (1986) on the other hand argued that error rate should decrease with the number of timesteps, citing two grounds for this claim. Firstly, he suggested that low numbers of timesteps would sometimes not allow enough time for the activation from a jolted morpheme to affect the activation of a target phoneme. This is true when the number of timesteps before selection is less than the number of layers between the morpheme layer and the phoneme layer, in which cases selection is entirely random. As noted in section 4.4.2 however, this does not apply to the current model, or many other models in the literature, as the word layer is directly connected to the phoneme layer, such that one timestep is sufficient for activation to be transmitted. Secondly, Dell (1986) argued that at low numbers of timesteps, there is not enough time for perseveratory activation from previous productions to decay, such that this activation interferes with the current production and causes errors. Whilst this argument did not apply to the productions on initial words that we considered when talking about first onset error rates, it does apply here.

Dell (1986) showed in a number of simulations that error rate decreases over time because perseveratory activation decreases. In a simulation of the SLIP task, a decrease in perseveratory activation drove a decrease in exchange errors as the number of timesteps used before selection was increased. Anticipations and perseverations were not affected. However, we do not consider this simulation further as the anticipatory and perseveratory bias activation which is directly applied to onsets at first and second onset production in order to mimic the SLIP procedure (a model feature not used anywhere else in the literature) makes the model behave quite differently to the general phonological encoding simulation presented by Dell (1986) upon which the current implementation is based. This difference is largely intentional, as the contextual error generation behaviour of the SLIP model was based

on results from the SLIP task, which differ quite greatly from the human corpus results which the general phonological model was based on. Correspondingly, at three timesteps, the SLIP simulation generates over six times as many exchanges as anticipations, whereas the general phonological encoding simulation generates at least three and a half times as many anticipations as exchanges. The SLIP simulation also hardly generates any perseverations, a behaviour which does not fit in with human behaviour in SLIP tasks and which Dell (1986) criticises.

Dell (1986) also provided evidence for the claim that error rate decreases over time in a simulation using the general phonological encoding model. We referred to this simulation when discussing first onset errors in section 4.4.2 and noted that it differs quite strongly from many other simulations in the literature as the activation levels of the representations in the model are not affected by activation-based or intrinsic noise. Any errors made by the model therefore represent a permanent inability to produce the specified phrase given the specified parameter settings.

We do not consider results when only two timesteps are allowed before selection. At this setting, errors are caused because it is impossible for activation from the jolted morpheme to travel across the two intermediary layers and arrive at the phoneme layer in the time allowed. Selection at the phoneme layer is therefore random. As previously explained, this is not a source of errors in our implementation or in many other models in the literature, because the word layer is directly connected to the phoneme layer, such that one timestep is sufficient for activation to reach the phonemes.

Instead, we consider the results reported for behaviour at three, four or five timesteps. Results are reported for productions of phrases of one word, two words, and six words. As these are all bisyllabic words, the results for production of phrases of one word, or two syllables, are perhaps most comparable with ours. Dell (1986) states that with three timesteps before selection, no errors occur on productions of two syllables. Similarly, across all specific models, we see an extremely low maximum error level of 1.48% at two timesteps. Our results show higher error levels at five and eight timesteps however, whereas Dell (1986) also reported no errors at four or five timesteps.

Error rates increase in Dell's (1986) simulation when longer phrases are produced. As no noise is present in the network and enough time is permitted for jolt activation to reach the phoneme level, the simulation only generates errors due to perseveratory activation, such that these errors must represent errors on non-initial onsets. For

these longer phrase lengths of two words (or four syllables) and six words (or twelve syllables), there is a very clear effect of timesteps on error rate, such that many more errors are generated at three timesteps (11.3% of phoneme productions for four syllables, 24.3% for twelve syllables) than at four timesteps (1.3% for four syllables, 5.7% for twelve syllables) or eight timesteps (0.5% for four syllables, 0.6% for twelve syllables).

It makes sense that contextual errors due to perseveratory activation should reduce as the number of timesteps per selection stage increases. Perseveratory activation will under most parameter settings be weaker when more timesteps are used, as it will have more opportunity to decay. However, on the first onset, our explanation of the increase of errors seen when higher steps settings were used focused on the idea that a higher number of timesteps causes the original jolt activation to decay and allows activation to spread to representations which are only distantly related, and thereby gives noise in the network more of an opportunity to distort the activation patterns and cause errors to occur. As there is no noise in the simulation of Dell (1986) which is summarised here, errors could not be generated in this way. More importantly, the lack of noise on representations means that the model differs crucially from our implementation, and many other implementations of Dell's (1986) theory (Dell & Gordon, 2003; Dell et al., 2004; Dell, Schwartz, et al., 1997; Hartsuiker, 2002; Foygel & Dell, 2000; Goldrick & Rapp, 2002; Martin et al., 1994; Rapp & Goldrick, 2000; Ruml & Caramazza, 2000; Ruml et al., 2000, 2005; Schwartz et al., 2006).

We therefore turn to the general simulation of phonological encoding reported by Dell (1986), where he uses exactly the same architecture and parameter settings as in the previously described simulation, but activation-based noise is applied to the activation levels of the representations. Here, Dell (1986) investigates the behaviour of the network when three, four and eight timesteps are allowed before selection. Dell (1986) does not refer to the results of this simulation as evidence for his claim that error rate decreases as the number of timesteps allowed before selection increases, but as this simulation is both a simulation of ordinary phonological encoding, rather than a simulation of a specific experimental task, and also involves noise being applied to the representations, it is useful for us to examine how its behaviour is affected by the number of timesteps which pass before selection occurs.

In this simulation, Dell (1986) only considers productions of pairs of bisyllabic words, such that four syllables are produced on each trial. The overall error rates

of the original simulations are not stated explicitly. However, the number of errors per error unit (phoneme, rime, cluster, consonant vowel pair, syllable) are provided, which is sufficient for us to approximate the error rates for each timestep setting.[2] Our calculations suggest that for three timesteps per encoding stage, 16.9% of phonemes were encoded incorrectly. This decreased as more timesteps were allowed, with an error rate of 5% at four timesteps, and 4.4% errors at eight timesteps.

Dell (1986) reports that all errors in his simulations were contextual errors. Using the breakdown of error types that Dell (1986) provided, we can roughly split errors into errors which are perseveratory and those which are anticipatory.[3] Those with a perseveratory component would be most in line with the second onset errors we consider here as they clearly occur on a non-initial word, and are at least partially due to influences of previous productions. Our calculations show that 11.9% of phoneme productions at three timesteps involved perseveratory errors, compared to 1.8% of productions at four timesteps and 0.6% of productions at eight timesteps. The remaining anticipatory errors accounted for 5.0% of the phoneme productions at three timesteps, 3.3% at four timesteps, and 3.8% at eight timesteps.

The perseveratory error rates are just slightly higher than the error rates that Dell (1986) reports that the model with no noise generates for productions of four syllables at three, four and five timesteps. As in the model with no noise, the perseveratory error rates decline rapidly as the number of steps before selection is increased. In line with this, perseveratory errors appear to be responsible for the overall decline in error rate as timesteps increase, whereas there is no clear trend of error rate change with timesteps in the remaining anticipatory errors.

There are three key differences between the results from Dell's (1986) simulation and our results. Firstly, Dell's (1986) results show that huge numbers of perseverations

[2]Dell (1986) reports that the simulations encoded 120 word pairs, where words were randomly selected from the model's entirely bisyllabic vocabulary, such that 480 syllables would have been encoded in total. The maximum length of a syllable was probably about 5 phonemes (two phonemes for an onset cluster, one phoneme for the nucleus, and two phonemes for a coda cluster). This means a rough maximum of 2400 phonemes were produced. Using the 5 phoneme assumption, the number of phonemes per error unit can be estimated (e.g., two phonemes for a phoneme cluster, and 5 phonemes for a syllable), and then by combining the counts of anticipations, perseverations, etc. per error unit (where exchanges and shifts are counted twice as they affect two units) we can roughly calculate the error rate for each timestep setting.

[3]When calculating perseveratory errors, we count the errors Dell (1986) classified as perseverations, perseveratory additions, exchanges, shifts, half the errors classified as ambiguous between anticipations and perseverations, and half the errors classified as deletions as these represent either anticipation or perseveration of a null segment. In contrast to the overall error calculation where exchanges and shifts are counted twice, exchanges and shifts are counted only once as only one part of the error will be perseveratory.

are generated when selection occurs after three timesteps, whereas in our results for second onset productions when selection occurs after two timesteps, error levels are really low, reaching a maximum of 1.48% across all models. Secondly, the error rates in Dell's (1986) simulations decrease as the number of timesteps increases, whereas error rates in our simulations increase as the number of timesteps increases. Thirdly, Dell (1986) reports that no non-contextual errors are generated, whereas the proportions of errors which are non-contextual on the second onset in our data are high, especially when selection occurs after five or eight steps. We argue that these differences are all related. The decrease in perseverations with increasing timesteps in Dell's (1986) results is largely driven by the very high number of perseverations generated when selection occurs after three timesteps. In contrast, the increase in error rates with increasing timesteps in our results is largely driven by non-contextual errors. We will examine these points in more detail individually.

Firstly, we consider the difference in the number of perseverations when a selection occurs after a low number of timesteps. This difference may be driven by the fact that our simulation concerns productions of two syllables, where as in Dell's (1986) simulation, four syllables are produced. There is actually no evidence that Dell's (1986) simulation generates any perseverations on productions of two syllables when a low number of timesteps is used. As previously noted, using the parameter settings that Dell (1986) chose, the model with no noise generates no errors at all on productions of two syllables. In contrast, when selection occurs after three timesteps, an extremely high error rate of over 11% is displayed in the model with no noise when four syllables are produced, and the error rate is over twice as high again when twelve syllables are produced. Dell (1986) does not test the behaviour of the model with noise on production of two syllables, but given that the perseveratory error rates for productions of four syllables are only slightly higher in the model with noise than in the model with no noise, it seems reasonable to assume that the same relationship would hold for productions of two syllables, such that the model with noise would also not generate many perseveratory errors on productions of two syllables.

The error rate exhibited by the model for productions of four syllables when selection occurs after three timesteps is very high, and we return to determining what might constitute an acceptable error rate in section 4.5. As the deterministic model without noise generates these errors, there is also an implication that Dell's (1986) model is unable to correctly produce quite a number of certain phrases at this

speed. However, we note for later references that in Dell's (1986) model, production of more syllables is likely to lead to an increase in the number of perseverations generated. It could be argued that an increase in perseveration errors may also be caused by not resetting the activation level of the target first word after first word phoneme selection. However, the lack of errors generated on productions of two syllables in Dell's (1986) simulations does not provide much evidence to suggest that this potential difference in implementations had a strong effect on the models' perseverative tendencies.

Secondly, we consider Dell's (1986) report that no non-contextual errors are generated in his simulation. Our results show that a large proportion of the extra errors that occur at higher timesteps settings are non-contextual. The number of contextual errors generated increases as the number of timesteps before selection is increased, with a median of 0.00% of onset productions resulting in a contextual error at two timesteps, 0.02% at five timesteps, and 1.59% at eight timesteps. However, the increase in the number of non-contextual errors generated is much larger, with a median of 0.00% of productions resulting in a non-contextual error at two timesteps, 0.03% at five timesteps, and 8.98% at eight timesteps.

We assume that greater numbers of non-contextual errors are generated when selection occurs after a higher number of timesteps, as more time passing will cause the original jolt activation to decay and allow activation to spread to representations which are only distantly related, thereby giving noise in the network more of an opportunity to distort the activation patterns and random phonemes to be selected. Whilst selection of random phonemes will generally result in non-contextual errors, the competing contextual phonemes should also sometimes be selected by random (cf. Vousden et al., 2000). It seems likely that some of the timestep related increase in contextual error productions in our results is due to the competing phoneme being selected in this manner, especially given the relative size of the increase in non-contextual errors. However, as there are 16 onset phonemes, and there are not 15 times as many non-contextual errors as contextual errors when selection occurs after eight timesteps, it seems that the contextual competing phoneme is still slightly more likely to be selected than non-contextual phonemes when the effects of noise on the network prevent the correct phoneme from being selected.

In a similar vein, it is possible that in Dell's (1986) simulation, some errors which were generated by noise mechanisms that usually produce non-contextual errors have been classified as contextual errors. In Dell's (1986) simulations, the phrase is twice as long as in our simulations, and so the set of phonemes whose production

would be classified as a contextual rather than non-contextual production is bigger. A larger group of contextual phonemes would also mean that there were more phonemes which were exposed to contextual activation, and our results suggest that these phonemes would be slightly more likely to be selected even when errors were driven by noise rather than extremely high amounts of perseveratory activation. An increase in these noise-driven errors as the number of timesteps before selection increases may be hidden by the huge swell of perseveratory activation driven errors when selection occurs after fewer timesteps.

It would otherwise be rather surprising that absolutely no non-contextual errors were generated in Dell's (1986) simulations. In implementations of Dell's (1986) theory which focus on single word production only, there is no context, and also no priming of upcoming words or perseverating activation from previous words. These models produce errors however, and these errors are therefore all non-contextual (Dell & Gordon, 2003; Dell et al., 2004; Dell, Schwartz, et al., 1997; Foygel & Dell, 2000; Goldrick, 2006; Goldrick & Rapp, 2002; Martin et al., 1994; Rapp & Goldrick, 2000; Ruml & Caramazza, 2000; Ruml et al., 2000; Ruml et al., 2005; Schwartz et al., 2006). Dell (1986) himself noted the lack of non-contextual errors generated by his simulation, and suggested that non-contextual errors could be generated by "increasing the background noise" (Dell, 1986, pp. 298), which fits in with the account we use here of the generation of non-contextual errors. However, our results clearly show that non-contextual errors can be generated with less activation-based noise than was present in Dell's (1986) simulations, and also with no intrinsic noise, as in Dell's (1986) original simulations, which suggests that this is not the main cause of the different behaviour reported in Dell's (1986) results.

Our results, combined with Dell's (1986) results, show that for the error rate to decrease as the number of timesteps before selection is increased, a model must show a strong tendency to generate perseverative errors when selection occurs after a low number of timesteps. Dell's (1986) results suggest that such a tendency is more likely to be exhibited when the model is producing phrases of four or more syllables. However, we emphasise that the perseverative error rates of 11% and 24% that Dell's (1986) model generates even without activation-based or intrinsic noise affecting the activation levels of representations are extremely high, and return to examining what sort of error rates would be acceptable in section 4.5. Furthermore, the fact that these errors are generated by the model without noise, which behaves entirely deterministically, suggests that Dell's (1986) theory implies that 11% of four syllable phrases and 24% of twelve syllable phrases are nearly impossible to

produce without sufficient time to prepare. It is not clear whether this conclusion is in line with human performance.

In models which don't show such a strong tendency to generate perseverative errors, such as those which produce shorter phrases, and in which noise affects the activation levels of representations, our results suggest that the error rate is likely to increase as the number of timesteps before selection increases. As previously argued when discussing the behaviour of the model on the first onset in section 4.4.2, we suggest that increasing the number of steps before selection increases the error rate as the jolt activation transmitted to the target phoneme has more time to decay, activation has more time to build on unrelated representations due to feedback loops, and noise has more time to affect the activation levels. This mechanism of error generation largely results in random errors, as demonstrated by the very high proportions of non-contextual errors when a selection occurs after a high number of steps. In these circumstances, we argue that the number of steps before selection is again better characterised as the length of time for which the model must remember the message it intends to convey, rather than speech rate or the time available to the model to prepare.

As a side note, we observe that it is surprising that Dell (1986) reports that no non-contextual errors are generated at all. This does not line up with our results or the high numbers of errors without a contextual source reported in models which focus on single word production. We suggest this result may be due to the larger context in Dell's (1986) simulations, such that there was a larger group of phonemes which would be activated by the context, and perhaps more importantly, whose errorful production would be classified as a contextual error.

Activation-based noise

The regression analyses shown in table 4.5 along with figure 4.5 show that higher values of *actiNoiseSD* lead to higher error rates. However, whereas *actiNoiseSD* was the second strongest predictor of first onset error rate, it has a below average effect on second onset error rate, with a Wald's Z value of 593.4, compared to a mean of 1234.3 for this measure. Higher values of *actiNoiseSD* also lead to higher proportions of non-contextual errors. This parameter is the second strongest predictor of second onset error non-contextuality, with a Z value of 104.8. However, this is very small in comparison with the Z value of the strongest predictor, *steps*, which is 477.1. It is also smaller than the Z value of the effect *actiNoiseSD* had on first onset non-contextuality, which was 133.6, where this parameter was the

weakest predictor of this measure. Correspondingly, the effect of *actiNoiseSD* on this measure is not very clearly visible in figure 4.6.

These results largely fit in with the effect of activation-based noise as explained for the first onset. Error rates increase at higher settings of activation-based noise as more noise means that the activation patterns caused by the jolt activation are distorted. The reduction in strength of this effect compared to behaviour on the first onset is probably due to the lack of a strongly primed competitor upon which the activation-based noise can act. The proportion of non-contextual errors increases as activation-based noise increases because the activation patterns left in the network by production of the first word are less clear, due to the effect of noise during both first and second word production. The stronger effect of activation-based noise compared to intrinsic noise is probably due to the fact that by second word production, some specific models will have accumulated a lot of activation on their representations which the activation-based noise can act upon.

Intrinsic noise

The logistic regression summarised in table 4.5 shows that higher values of *intrinNoiseSD* lead to higher second onset error rates, as they did on the first onset. As was also the case for the first onset error rate, this is the second weakest effect on this measure, with a Wald's Z value of 300.2, compared to a mean of 1234.3. Similarly, higher values of *intrinNoiseSD* lead to higher proportions of non-contextual errors, but this is the weakest effect of all, with a Z value of 48.3, compared to a mean of 146.0. On the first onset, two other parameters had less of an effect on error non-contextuality. Both of these effects are visible in figures 4.5 and 4.6, but are also visibly small.

As on the first onset, an increase in intrinsic noise reduces the influence of the jolt activation, which leads to more errors being produced. Such an increase also distorts the activation from previous productions, resulting in lower proportions of contextual errors, or higher proportions of non-contextual errors. However, the relatively weak effects suggest that the aberrations in activation levels caused by intrinsic noise are small in comparison to the manipulations of activation levels effected by most other parameters. This may be particularly true on the second onset, where representations in some specific models may have accumulated large amounts of activation, in comparison to which the intrinsic noise dwindles.

4.4.4 Summary of effects of parameter manipulations on first and second onset behaviour

In the previous two sections, we saw that some parameters have roughly the same effect on both the first and the second onset. These are connection strength, the number of steps, and the amount of activation-based and intrinsic noise.

It was shown that problems in processing can be caused if connection strength is either too low or too high. At connection strengths which are too low, activation cannot be effectively transmitted from the jolted word, causing more errors to be generated (c.f. Dell, Schwartz, et al., 1997). Activation from the primed word at the first onset also has less influence, resulting in fewer errors. On the second onset, contextual errors are also reduced as the first onset was never sufficiently activated to cause them. At connection strengths which are too high, representations in the network become inappropriately highly activated. We argue that this is due to too strong an influence of feedback (c.f. Goldrick, 2006), either due to strong feedback connectivity, or strong forward connectivity strengthening the forward flowing part of the feedback loop. On both the first and second onset, this causes the network to begin to behave more randomly, leading to an increase in error rate and non-contextuality of errors.

Our results showed that higher numbers of timesteps before selection lead to higher error rates and higher proportions of non-contextual errors on both the first and second onset. When too long a period of time passes between the jolt activation being applied to the target word and selection at the phoneme level, the network is unable to retain its intended message and errors occur. On the first onset, errors become more non-contextual because the prime of the upcoming word also becomes less effective. On the second onset, increasing numbers of steps mean that the jolt of activation to the first onset is further and further in the past, leading to greater proportions of non-contextual errors. We highlight that this result is contrary to Dell's (1986) claim that the accuracy of the network increases as the number of timesteps between jolt and selection increase, and causes problems for the links he drew between a slow speech rate (at which humans make fewer errors; MacKay, 1971) and a high number of steps being allowed before selection. Our results show that at higher timesteps settings, the decay of the original signal combined with the increased influence of noise on the activation levels and the opportunity for activation to build on unrelated representations via feedback loops appears to make the network behave very randomly. On the second onset, higher error rates may

be expected at lower timestep settings due to stronger perseverative activation, but our results and Dell's (1986) results suggest that for productions of two syllable phrases, the perseverative activation which accumulates is not strong enough to outweigh the effect of accumulating noise. We therefore suggest that, particularly in single word production models, the number of steps before selection is generally better characterised as the amount of time for which the network has to retain the intended message.

It was also shown that increasing activation-based noise or intrinsic noise increases error rates and non-contextuality of errors on both onsets. In both cases, increased noise makes the activation signal from the jolt less distinct, leading to more errors. Noise also reduces the influence of the prime on the first onset, and muddies the activation patterns from the first onset production, such that they have a less distinct effect on the second onset, thereby causing a greater proportion of non-contextual errors to be generated.

The remaining parameters, decay, and jolt and prime, have different effects on the two onsets.

On the first onset, a high rate of decay causes the jolt and prime activation to fade away, leading to a higher error rate and higher proportion of non-contextual errors (c.f. Dell, Schwartz, et al., 1997). On the second onset, a high decay rate serves to clean the network and leave it freer of interference from the production of the first onset, resulting in fewer errors. However, as the previous production has a lesser influence at higher decay rates, a lower proportion of contextual errors are generated (c.f. Dell, 1986).

As prime activation is a concept which only makes sense when considered relative to jolt activation, we discuss the effects of these two parameters together. On the first onset, a prime which is high relative to the jolt increases the likelihood that the upcoming onset will be anticipated (c.f. Dell, 1986), leading to higher error rates and lower proportions of non-contextual errors. Another way of looking at this result is that the network is more accurate when the jolt applied to the current word is stronger in comparison to the rest of the activation in the network (c.f. Goldrick, 2006). On the second onset, a high jolt again reduces error rate because the activation from the current word is stronger. A high prime increases error rate because it causes there to be more activation in the network for the jolt to compete with, and it also increases the chance that there was an error on the previous production such that the activation on the first onset has not been

suppressed. Overall therefore, a high jolt to prime ratio leads to a lower error rate. The effect of jolt and prime on non-contextuality of second onset errors is a little more complicated. We suggest that a low jolt to prime ratio increases the likelihood of exchange generation. Relatively higher primes make the anticipatory portion of the error more likely, causing the activation on the first onset not to be suppressed. The second onset, whose activation has been suppressed, then receives relatively less jolt reactivation, making it more likely that the first onset will be selected in error. A high jolt to prime ratio on the other hand makes production of perseverations more likely. Following correct production of the first onset, a relatively lower prime means that the non-suppressed second onset has received less preactivation. At the same time, more activation has been passed to the neighbours of the first onset, making it more likely that they are able to reactivate the first onset to a level at which it is reselected in error. Together, these effects mean that the tendency for contextual errors to be produced is lowest at mid-range jolt to prime ratios. We will further examine these claims about the effect of jolt to prime ratio on perseverations and exchanges in the following chapter.

Some parameters had universally strong effects at the settings we tested, whereas others were generally quite weak. Specifically, the effect of the number of timesteps was very strong for all measures. At two timesteps, the model was very well behaved. On both onsets, hardly any errors were generated in any specific models, and very few specific models generated more than a very low proportion of non-contextual errors. At eight timesteps however, many specific models experience problems recalling the intended message and the effects of primes and activation levels at previous productions have also faded away, leading to great tendencies to generate high numbers of errors, and very high proportions of non-contextual errors too, especially on the second onset.

Conversely, the effect of intrinsic noise at the settings we tested was generally small. Clearly, the blurring of activation levels caused by intrinsic noise at the range of levels we chose is weak in comparison to noise which is boosted by the activation levels themselves, or the manipulations of activation levels which result from changing other parameters.

It was also clear that certain parameters were particularly influential on specific measures. Beyond the universal effect of steps, activation-based noise, jolt to prime ratio and also connection strength were key in determining the first onset error rate of a specific model. Activation-based noise and the jolt to prime ratio were particularly important because of the presence of the primed competitor. While the

jolt to prime ratio determined how strong the prime was, the activation-based noise particularly affected the activation levels of the highly activated primed and target representations. Our results show that inappropriately high activation of representations due to high connection strengths was also an important cause of errors. Again, on top of the effect of the number of steps per stage, the jolt to prime ratio was important in determining what proportion of errors were non-contextual on the first onset, whereas the level of activation-based noise had a particularly weak effect. The importance of the jolt to prime ratio is easily explained, because contextual errors are only generated on the first onset when the prime is high. The effect of activation-based noise was muted as this noise component has a big effect on the activation level of the one primed phoneme, whose production results in a contextual error, and a small effect on the activation level of the many other phonemes, any of whose production would result in non-contextual errors. These two effects balance off to result in similar proportions of contextual and non-contextual errors regardless of the level of activation-based noise.

On the second onset, connection strength and the decay rate are important predictors of error rate along with the number of timesteps per stage, whilst there is very little effect of jolt to prime ratio. We suggest that high connection strength is a particular cause of errors here because it causes activation levels to really rise during production of the first onset, and on the second onset has further opportunity to boost the already high levels of activation. Decay becomes important because it affects how much irrelevant activation persists from the previous production. The jolt to prime ratio is less influential here as no representations are directly primed, and the activation scale of the network has already been set. Non-contextuality of second onset errors is nearly entirely decided by the number of timesteps per selection stage. Our results show that the number of steps which have passed since the first onset jolt, varying from 4 to 16 in our settings, is by far the most important determiner of how much influence previous productions have on the current production.

4.5 Limits on error rate and non-contextuality of errors

The previous two sections have demonstrated that certain parameter settings of the spreading activation model lead it to generate very high numbers of errors, or very high proportions of non-contextual errors, or both. Researchers such as Dell, Schwartz, et al. (1997), Foygel and Dell (2000) and Rapp and Goldrick (2000) have relied on these manipulations to capture patterns of aphasic error behaviour.

However, the present thesis focuses on modelling data from the normal population. This section first uses experimental and corpus data to establish upper limits on error rates and proportions of non-contextual errors produced by normal speakers, and then examines which specific models generated acceptable error rates and appropriate proportions of contextual errors according to our newly established limits, which specific models generated either too many errors or too high a proportion of non-contextual errors, and which specific models did not generate any errors at all.

4.5.1 Establishing limits from human performance data

In determining bounds on the error rates and proportions of non-contextual errors generated by normal speakers, we erred on the side of liberal overestimates of these measures in order to consider the biggest possible set of acceptable parameter settings. Settings can later be pruned from this set using more conservative constraints if required.

It is difficult to use speech error corpora to create bounds for overall error rate, as nearly all such corpora consist solely of a collection of errors noted down by the investigator as they occur, and do not hold the information necessary to determine what proportion of speech these errors account for. To approximate a liberal upper bound on error rate, we therefore referred to one of the primary sources of experimental data for this thesis, Goldrick and Blumstein (2006). This is the only experiment modelled in this thesis where an error rate is reported, and the errors referred to are those fed into the analysis which we model. We examined the range of onset error rates reported across participants, which was 0.5% to 4%, and doubled that range around its midpoint, such that the lower limit was floored at 0%, and the upper limit on the number of errors produced was 5.75%. This limit should be particularly liberal for simulation of normal phrase production, as tongue twisters are specifically intended to cause errors. For comparison, Dell, Schwartz, et al. (1997) reported a 3.1% error rate on a picture naming task for normal control participants, including 0.7% non-naming responses such as descriptions, and 2.1% of productions which comprised errors with a clear semantic relation to the target, whereas the implementation described in this thesis can only generate errors at the phoneme level or below. An analysis by Schwartz et al. (1994) of the London-Lund corpus (Garnham et al., 1981), the one speech error corpus we know of for which a full record of non-erroneous speech also exists, revealed an overall error rate of 0.07%, including 0.03% sound errors (as opposed to errors involving grammatical elements, words or larger units), which is substantially lower than our upper limit.

Determining limits on the non-contextuality of speech errors does not require information about non-erroneous speech, and so error corpora which report numbers of both contextual and non-contextual errors can be used. We found four corpora which contained this information (del Viso et al., 1991; Pérez et al., 2007; Shattuck-Hufnagel & Klatt, 1979; Vousden et al., 2000). The proportions of non-contextual errors found in the error corpora and further detail on the corpora are given in table 4.6.

When analysing the corpora, we considered the fact that our simulation analyses would be based on within-clause onset consonant substitution errors only, and used the corpus figures provided for the most similar subset of errors. Table 4.6 describes exactly what this subset was on each occasion. To calculate the proportion of non-contextual errors in each corpus, we counted up anticipations and perseverations, and assumed that a complete exchange was worth two contextual errors, as the lack of an identifiable source presumably meant that each non-contextual error only involved one error location. All incomplete errors such as *"big fun"* → *"fig. . . "* were classified as one contextual error for this analysis.

The proportion of non-contextual errors determined from Vousden et al. (2000) may overestimate non-contextual substitutions in comparison to other analyses. Vousden et al. (2000) attempted to compensate for the fact that random phoneme substitutions will sometimes resemble a contextual phoneme substitution because they happen to result in the production of a phoneme which is in the context. They calculated how often chance predicts that this should happen, and moved a corresponding portion of the contextual substitutions recorded in the corpora to the non-contextual substitution category. However, as we aimed to calculate liberal overestimated limits on non-contextuality, and the resulting proportion of non-contextual errors (15.9%) was still within the bounds of proportions found in other corpora, we judged it reasonable to include these figures in our analyses.

Of the four corpora in our analysis, the two highest proportions of non-contextual errors came from the English corpora, and the two lowest proportions came from the Spanish corpora. However, with so little data available, we considered data from both languages as representative of normal speaker behaviour for these analyses.

To determine bounds on non-contextuality, we calculated the mean proportion of non-contextual errors across the corpora, which was 14.75%, and set the liberal limit at two standard deviations around the mean. With a standard deviation of

8.64%, the lower limit was again floored at 0% of errors being non-contextual, and the upper limit set at 32.04%.

4.5.2 Which specific models met the limits on error rate and non-contextuality?

To determine which specific models exhibited behaviour which was within the limits on error rate and non-contextuality of errors set out in section 4.5, we calculated the error rate and non-contextuality of errors for both onsets combined. Over all 5832 specific models, we found that 1727 specific models generated no errors, which was 29.6% of those tested. Another 2440 specific models, which was 41.8% of those tested, failed either the constraint on error rate or the constraint on non-contextuality of errors, or both. Of these 2440 specific models, 1465 generated both too many errors and too high a proportion of non-contextual errors (25.1% of all models), a further 504 generated too many errors despite producing a sufficiently low proportion of non-contextual errors (8.6% of all models) and the final 471 generated sufficiently few errors, but too high a proportion of non-contextual errors (8.1% of all models). The remaining 1665 specific models, representing 28.5% of the models tested, generated errors but succeeded in exhibiting a sufficiently low error rate and producing a low enough proportion of non-contextual errors to pass our constraints.

The effect of parameters on the combined onset error rate and non-contextuality of errors is shown by the regression summary in table 4.7, and in figures 4.7 and 4.8. Building on this, figure 4.9 shows how many specific models per parameter setting generated no errors, how many generated too many errors, how many generated too high a proportion of non-contextual errors, how many failed both the constraints on error rate and non-contextuality of errors, and how many passed both of these constraints.

In nearly all graphs in figure 4.9, the same number of specific models is tested at all parameter settings. The dependent variable is therefore the number of specific models which fall into each category as set out in the previous paragraph. The exception to this rule is the *joltPrimeRatio* parameter. As explained in section 4.4.1, there are twice as many specific models with a *joltPrimeRatio* of 2 than there are at other *joltPrimeRatio* settings. To facilitate comparison between the behaviour of models at different *joltPrimeRatio* settings, the dependent variable in the *joltPrimeRatio* graph is therefore the percentage of specific models which fall into each of the specified categories.

Table 4.6: Speech error corpora used for analysis of non-contextual error proportions.

Name	Source paper	Collectors	Language	Number of errors in corpus	Errors considered for analysis	Number of errors in analysis[a]	Number of contextual errors[b]	Number of non-contextual errors	Proportion of non-contextual errors
Del Viso et al.	del Viso et al. (1991)	Two highly theoretically informed trained observers [c]	Spanish	3,611	Phoneme substitution errors	476	489	71	12.7%
MIT-CU	Shattuck-Hufnagel and Klatt (1979)	Shattuck-Hufnagel and Garrett	English	6,000	Consonant substitution errors	1,620	1,409	486	25.6%
PSP	Pérez et al. (2007)	737 students on Psychology of Language course at University of Granada	Spanish	7,480	Phoneme substitution errors	979	1,142	57	4.8%
Vousden et al.	Vousden et al. (2000)	Harley	English	2,289	Phoneme substitution errors	2,251	2,096	397	15.9%

[a] In this category, exchange errors are counted as one error.
[b] In this category, exchange errors are counted as two errors, for reasons explained in the text.
[c] According to Pérez et al. (2007).

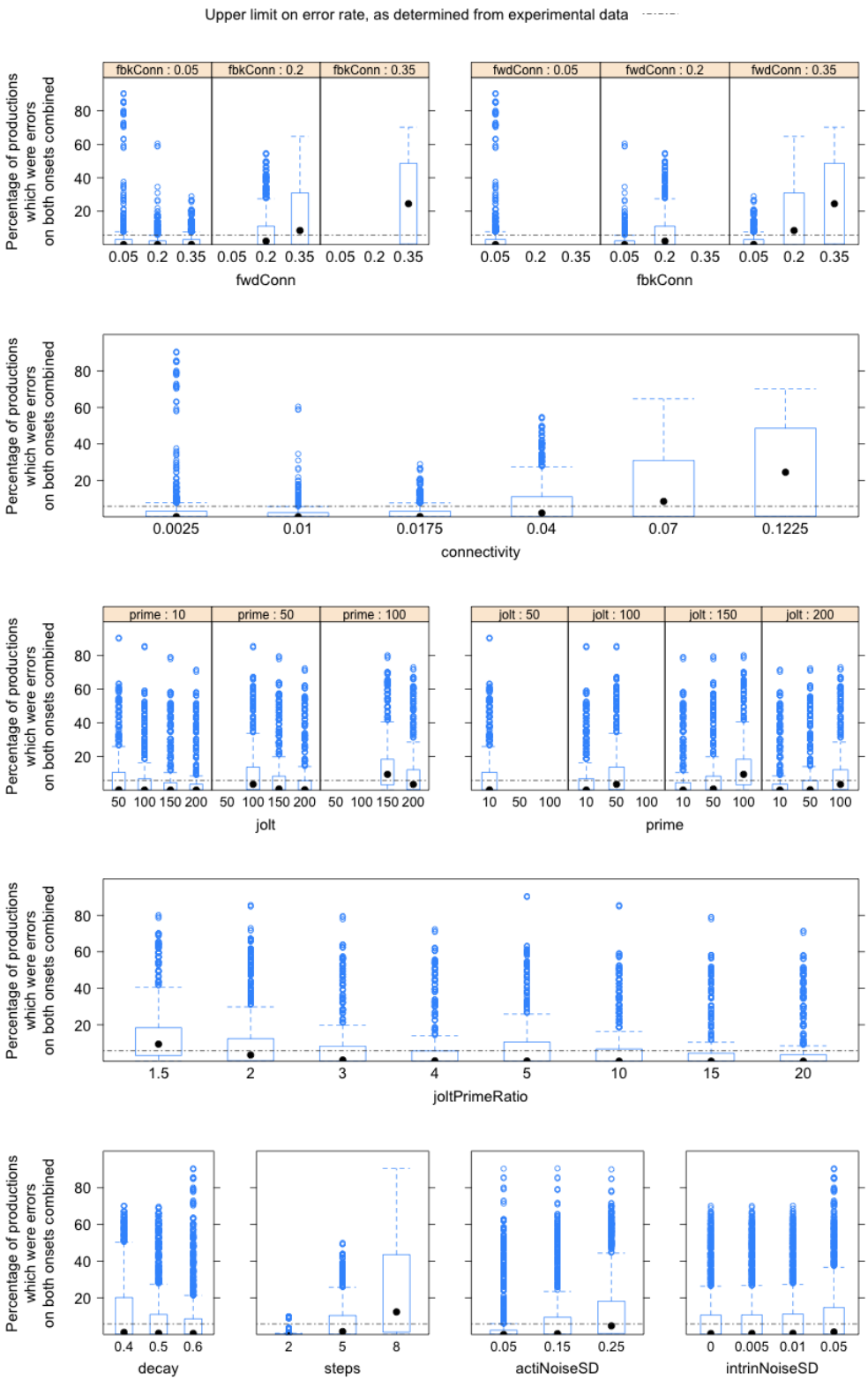


Figure 4.7: The effect of changing parameter settings on the onset error rate for both onsets combined. The dotted line represents the upper limit on error rate.

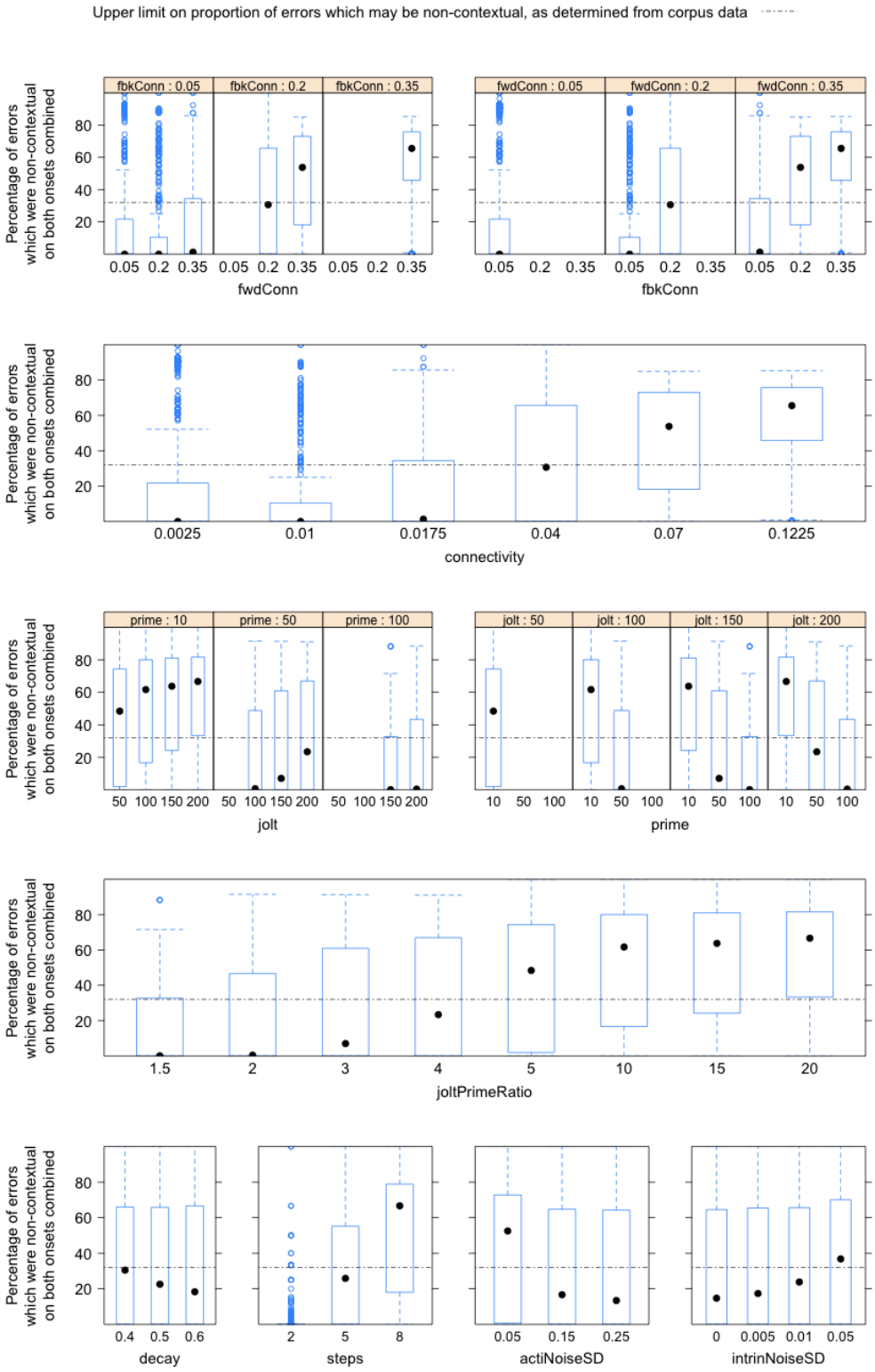


Figure 4.8: The effect of changing parameter settings on the proportion of errors which are non-contextual at both onsets combined. This proportion can only be calculated for specific models which generated at least one error. The dotted line represents the upper limit on error non-contextuality.

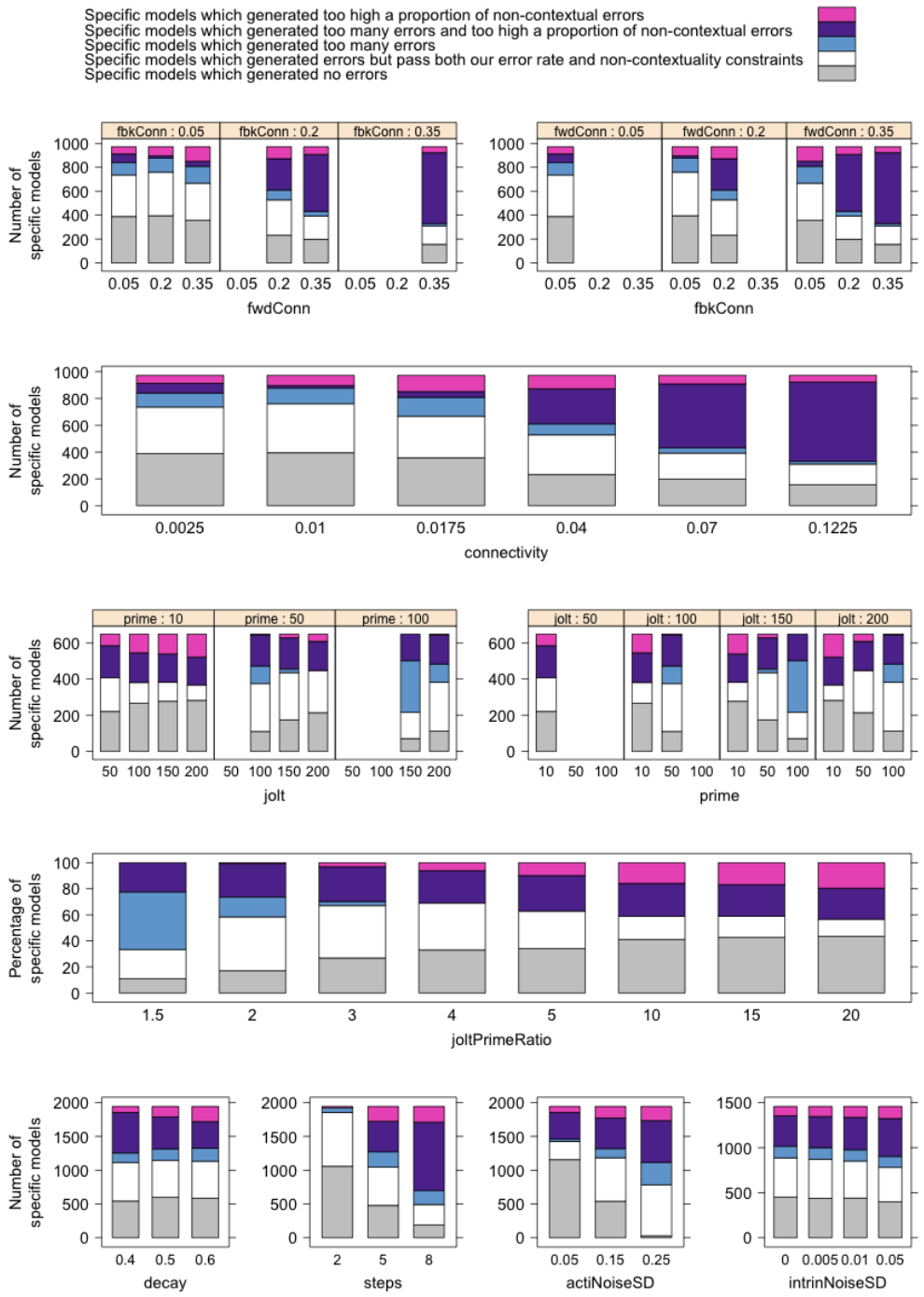


Figure 4.9: The effect of changing parameter settings on the numbers of specific models which pass our constraints, for all specific models. From the bottom of a bar to the top, specific models are classified as specific models which did not generate any errors; specific models which generated some errors, but not too many, and not too high a proportion of non-contextual errors (passing both of our constraints); specific models which generated too many errors overall, although an acceptable proportion of non-contextual errors (failing one of our constraints); specific models which generated too many errors overall, in addition to which too high a proportion of the errors were non-contextual (failing both of our constraints); and specific models which generated an acceptable number of errors overall, but too high a proportion of non-contextual errors within them (failing one of our constraints).

Table 4.7: Results of logistic regression model analyses using parameter values to predict error rate and proportion of errors which were non-contextual on both onsets combined. The proportion of errors which were non-contextual can only be calculated for specific models which generated at least one error. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	**Error rate**					**Non-contextuality**				
	Dir	Z	LRT	P (χ^2)		Dir	Z	LRT	P (χ^2)	
connectivity	+	2912.5	9669723	< .001	*	+	675.3	477192	< .001	*
joltPrimeRatio	−	737.9	585501	< .001	*	+	553.2	343102	< .001	*
decay	−	784.8	628174	< .001	*	−	84.8	7195	< .001	*
steps	+	2864.9	12877897	< .001	*	+	887.9	890592	< .001	*
actiNoiseSD	+	1084.1	1221311	< .001	*	+	13.9	192	< .001	*
intrinNoiseSD	+	391.4	150532	< .001	*	+	252.7	65066	< .001	*

Key:
Dir = direction

In line with the effects we observed on the first and second onsets individually, our analyses show that specific models with higher connectivity strengths and higher numbers of steps per selection stage frequently exhibit error rates and non-contextuality proportions which are too high to pass the constraints. The number of specific models which fail the constraints is higher for specific models with the lowest forward and feedback connection strengths than for specific models with slightly higher forward connection strength however. Specific models with lower connectivity strengths and lower numbers of steps often do not generate any errors for analysis.

Specific models with high activation-based noise settings are more likely to have inappropriately high error rates and generate too high a proportion of non-contextual errors, but correspondingly less likely to have no errors for analysis. The relative weakness of the effect of manipulating intrinsic noise is also depicted in figure 4.9, where it is clear that both the number of specific models generating any errors and the number of models failing the constraints on error rate and non-contextuality of errors only increases slightly at higher intrinsic noise settings.

The effect of altering the jolt to prime ratio largely reflects the strong effect of the parameter on error production on the first onset. Specific models with low jolt to prime ratios (i.e., with prime settings which are big in comparison to the jolt settings) often generate too many errors to be acceptable, whereas simulations with

higher jolt to prime ratios frequently suffer from either not generating any errors for analysis, or generating too high a proportion of non-contextual errors. Manipulating the decay rate has a reasonably muted effect on the final classification of simulations however. As shown in figure 4.7 and the analysis summarised in table 4.7, the error rate for both onsets combined increases as decay rate increases, reflecting the strong rise in second onset errors for higher decay settings. Figure 4.8 and the analysis in table 4.7 further show that a decrease in decay results in a decrease in the proportion of errors which are non-contextual across both onsets combined, probably because there are fewer second onset errors at higher decay rates and these have a much greater tendency to be non-contextual than first onset errors do. However, figure 4.9 implies that these effects do not result in a strong change in the number of specific models able to generate errors, or the number of specific models generating too many errors or too high a proportion of non-contextual errors to pass the constraints specified earlier.

4.6 Effects of parameter manipulations on the number of productions aborted due to zero selections

As explained in section 3.4.4, simulations occasionally encountered the problem that all the nodes in a phoneme group where selection was to occur had no activation, such that if selection proceeded, a node would be selected at random. We assume that on such occasions, the human word production system would abandon the utterance, and in line with this assumption, some productions are aborted due to what we refer to as *zero selections*. In the final analysis of this chapter, we examine how often zero selections occurred and which specific models were more likely to suffer from this problem.

Our results show that zero selections are very rare. Of the 5832 specific models tested, 5785 specific models, or 99.2% of the models tested, do not abort any productions. Across the remaining 47 specific models (0.8% of the models tested), the median number of zero selections aborted in a single specific model was 3, constituting 0.03% of the total 10,000 productions made by that model, and the maximum number of zero selections was only 18, representing 0.18% of the total productions.

These results imply that aborting productions due to zero selections will not have strongly affected our results. However, in the same way that the parameter manipulations affected the error rate and proportion of non-contextual errors generated by a specific model, they also affect how likely it is that zero selections occur. We use

the same graph and logistic regression approach which has been applied throughout this chapter to examine which specific models display greater tendencies towards zero selections.

Figure 4.10 and the logistic regression summarised in table 4.8 show that specific models with low connection strengths, high decay rates, a high number of steps before selection, and larger amounts of intrinsic noise are more likely to abort productions due to zero selections. The regression in table 4.8 shows that the effect of the jolt to prime ratio is not significant, but figure 4.10 gives some hint that the absolute sizes of the jolt and primes rather than the relative sizes may affect this behaviour, such that lower amounts of jolt and prime make a specific model slightly more prone to aborting productions through zero selections. Both the graph in figure 4.10 and the regression in table 4.8 show that activation-based noise has little effect on the likelihood that a specific model will abort productions due to zero selections.

We argue that most of the factors shown to increase the probability that a specific model will abort productions due to zero selections do so because they reduce how much activation there is in the network, making it more likely that a group of nodes will have no activation. Specifically, low connection strengths mean that less activation is transmitted from node to node; high decay rates cause activation levels to decay more quickly; higher numbers of timesteps before selection mean that the activation has more time to decay; and lower jolt and prime values mean that less activation is input into the network to begin with. It also follows that intrinsic noise has a greater effect on specific models' tendencies towards this behaviour than activation-based noise does, and that more intrinsic noise makes zero selections more likely. If the activation level of a node is very low, then the amount of activation-based noise which acts on that node will also be very low. In contrast, intrinsic noise does not decrease in proportion to activation, and higher amounts of intrinsic noise may therefore prove sufficient to reduce low activation levels to zero.

Whilst we have shown that specific models with low connection strengths, high decay rates, high numbers of steps before selection, high amounts of intrinsic noise, and low jolt and prime values are more likely to abort productions due to zero selections, these results generally demonstrate that even in models with the greatest tendency to abort simulations in this way, the problem affects an extremely low percentage of productions. Exclusion of this small number of productions from analysis is therefore unlikely to have any great effect on subsequent results.

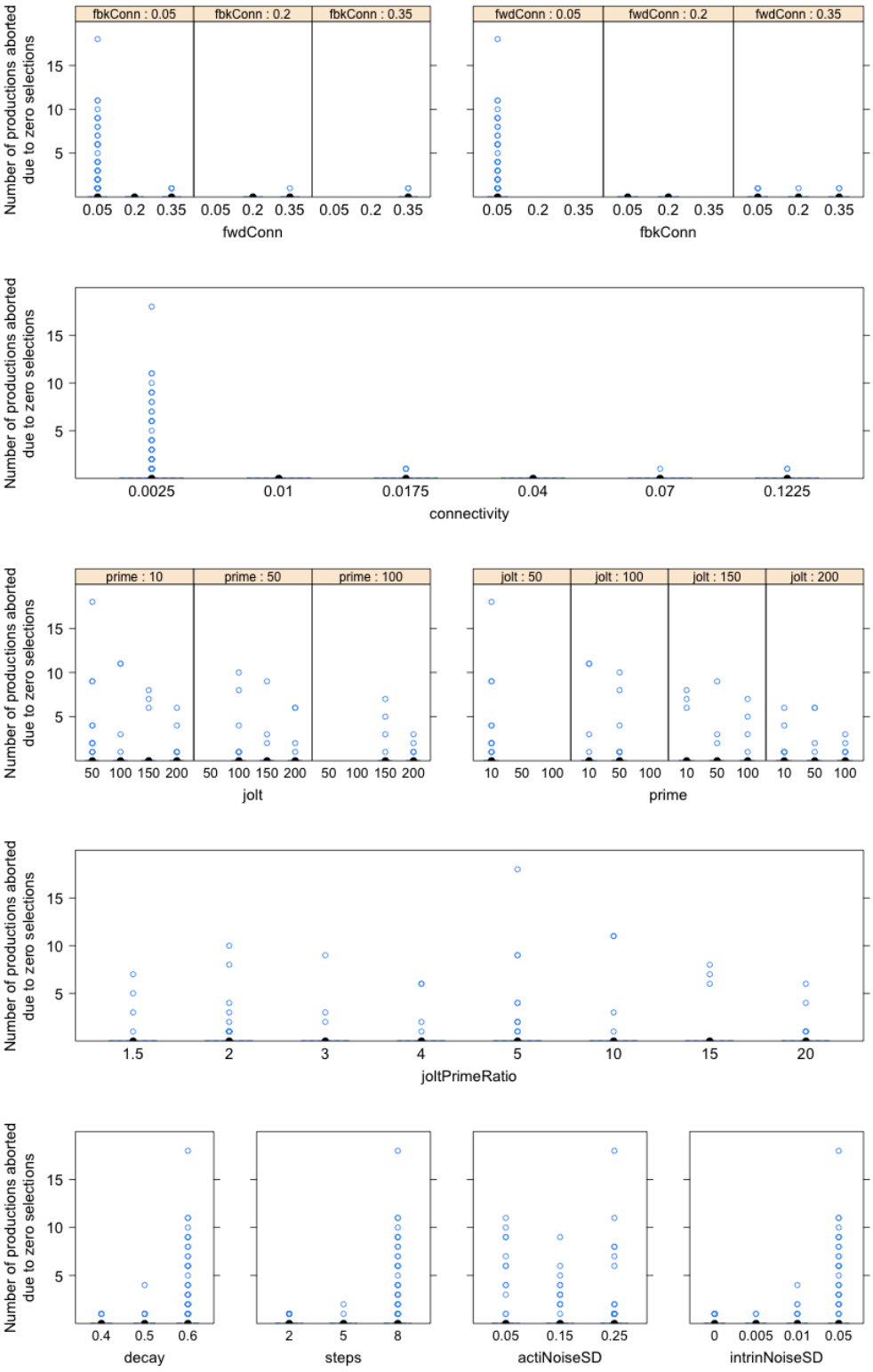


Figure 4.10: The effect of changing parameter settings on the number of phrase productions aborted due to zero selections, out of a total of 10,000 attempted productions for each specific model.

Table 4.8: Results of logistic regression model analyses using parameter values to predict how many productions are aborted due to zero selections. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	−	11.0	489.1	< .001	*
joltPrimeRatio	−	0.3	0.1	0.733	
decay	+	10.6	325.1	< .001	*
steps	+	10.7	319.2	< .001	*
actiNoiseSD	+	0.6	0.1	0.806	
intrinNoiseSD	+	13.8	364.5	< .001	*

4.7 Conclusions

In this chapter, we first looked at the error rate and proportion of errors which were non-contextual on the first and second onsets individually. Initial observations showed that there was great variation in these measures across simulations, confirming that manipulating the parameter settings had had important consequences for the behaviour of the implementation. The average error rate did not differ greatly between the two onsets, with medians between 0 and 1% in both cases. However, whilst the median proportion of errors which were non-contextual was very low on the first onset, it was very high on the second. Finding that the model had a strong tendency to generate non-contextual errors on the second onset was not in line with results from Dell's (1986) original simulations, where Dell (1986) reported that no non-contextual errors were generated. However, in section 4.4.3, we argued that this difference in behaviour was most likely due to the longer phrases of four syllables produced in Dell's (1986) simulations, compared to the phrases of two syllables produced in our own simulations. Longer phrases create a bigger context and a greater set of phonemes whose production would be classified as a contextual rather than non-contextual error.

We then introduced a methodology to allow us to obtain an overview of the effects of parameter manipulations on the implementation's behaviour, based on a regression approach which required us to transform four of our spreading activation parameters, and graphs of a specific and consistent format. This methodology is used throughout the thesis. We applied the methodology in this chapter to create a reference directory of the effects of parameters on the basic behaviour of the

network on first and second onset productions, and linked these results back to previous investigations distributed throughout the literature which have demonstrated some effects of manipulating individual parameters, as summarised in section 2.4.1. Later in the thesis, we refer back to the findings described here when interpreting effects of parameters on more complex behaviour.

As well as being of use in helping us interpret these more complex results, these investigations into the effects of parameter manipulations on error rate and non-contextuality of errors revealed some notable results of their own. Firstly, we highlight that for the network to avoid generating too many errors and too high a proportion of non-contextual errors, forward and feedback connection strength must be neither too low nor too high. At low connection strengths, activation cannot be transmitted effectively, leading the network to behave randomly (cf. Dell, Schwartz, et al., 1997). At high connection strengths, the effect of feedback becomes too strong (cf. Goldrick, 2006; Shrager et al., 1987), leading to inappropriately high activation of representations in the network, such that the activation levels of the target and primed phonemes are less distinguishable. This effect can be amplified either by increasing the feedback connection strength itself, or by increasing the forward connection strength, which boosts the forward streaming part of the feedback loop.

Secondly, we note that there were signs of a potentially interesting effect of the jolt to prime ratio on contextual error generation on the second onset. We suggest that a low jolt to prime ratio increases the probability of contextual second onset errors being generated as part of an exchange, but a high jolt to prime ratio increases the probability that second onset errors will be generated as part of a perseveration. We argue that this corresponds to the two different mechanisms proposed by Dell (1986) for the generation of second onset errors as part of perseverations and exchanges. This suggestion is examined further in the next chapter.

Thirdly, we highlight the different effect that decay rate has on error rates on the first and the second onset. On the first onset, increasing the decay rate increases the error rate, because it increases the tendency of the network to lose track of the intended production (cf. Dell, Schwartz, et al., 1997). On the second onset however, increasing the decay rate decreases the error rate, because higher decay results in more effective purging of activation from previous productions (cf. Dell, 1986).

Fourthly, and most significantly, we observe that higher numbers of timesteps before selection result in higher error rates and higher proportions of non-contextual errors on both the first and second onset. This is because higher number of steps provide

a greater amount of time for jolt and contextual activation to decay, for activation to spread to only distantly related representations, and for the effect of noise on activation patterns in the network to build up.

This result is in direct contrast to Dell's (1986) claims about the effect of manipulating the steps parameter, in which he argued that higher numbers of steps before selection increased the accuracy of the network, and were akin to slower speech rates. We note that Dell's (1986) claim holds in two cases only. In the first case, the number of timesteps before selection must be less than or equal to than the number of layers between the jolted node and the layer at which selection will occur. As this configuration prevents any jolt activation reaching the selection layer, selection is random. As soon as the number of timesteps is greater than the number of layers between the jolted node and the selection layer, performance improves dramatically as selection is no longer random. This situation does not apply to most models in the literature however, as in these models, layers at which jolt activation is applied are directly connected to layers at which selection occurs, such that one timesteps is sufficient for activation to be transmitted. In the second case, activation perseverating from previous productions is extremely strong. The decay of this activation with time and reduction of the associated errors outweighs the effect of decay of the jolt activation, activation spreading throughout the network, and noise, and the increase in errors that these factors lead to. However, our results alongside Dell's (1986) results suggest that perseverative activation only builds in this way when strings of more than two syllables are produced, unlike in our investigations. In Dell's (1986) investigations, when activation does build in this way, really high levels of perseverations are generated when selection occurs after a low number of timesteps, with over 11% errors on productions of four syllables, and over 24% errors on productions of twelve syllables. This is even without activation levels of the network being affected by noise, and these error rates are well above the liberal limit of 5.75% errors which we established from human tongue twister data in section 4.5. More notably, as these results occur in a deterministic model which is unaffected by noise, these results suggest that for certain phonemes in certain phrases, the perseverating activation is normally higher than the activation on the target phoneme, and the model is simply unable to produce these phrases when so few timesteps pass before selection. It remains to be seen whether this behaviour is in line with normal human performance.

Our results show that on networks which are not tested when i) the number of steps before selection is less than or equal to the number of layers separating the layer

at which jolt activation is applied from the layer at which selection occurs; and ii) in which perseverative activation does not build to such extreme levels; and finally, iii) in which noise is present, a higher number of timesteps before selection leads to a stronger hold of noise on the activation patterns in the network. Aided by the decay of the original jolt activation and the spreading of activation throughout the network, this results in higher error rates as well as higher proportions of non-contextual errors. We argue that in this common situation, rather than conceptualising the number of steps before selection as the inverse of speech rate (Dell, 1986), so that higher numbers of steps correspond to a slower speech rate, the effect of manipulating this parameter suggests that it is closer to a representation of how long the network has to remember what it is saying.

After examining the effects of the parameters on the basic behaviour of the network, we turned to experimental and corpus data to establish what sort of error rate and proportion of non-contextual errors would be in line with normal speaker performance. We determined liberal upper limits of a 5.75% error rate, and 32.04% of errors being non-contextual. The behaviour of our simulations was then compared to these constraints. Just under 30% of the simulations exhibited behaviour which passed both of these limits, whilst over 40% were too erroneous, with approximately even numbers of simulations failing the error rate and non-contextuality constraints, and 25% of all simulations failing both constraints. Nearly 30% of simulations did not generate any errors at all. We then outlined which specific models fell into each of these categories. The results which followed naturally from our previous investigations into the effect of parameter manipulations on error rate and non-contextuality of errors. These classifications are also used later in the thesis, allowing us to verify that claims that a model class can capture certain behaviour patterns are based on specific models which display basic error generation behaviour appropriate to normal speakers.

Finally, we showed that over 99% of the specific models that we tested do not abort productions due to zero selections. The extremely small number of specific models which do are generally models in which activation levels are low. We showed that this is due to low connection strengths, high decay rates, high numbers of steps, and low jolt and prime sizes, and where these low activation levels are further aggravated by higher intrinsic noise. Even these specific models abort very few productions for this reason however, with the highest abortion rate at 0.18% productions. These results strongly imply that excluding productions from analysis when zero selection

occur will have very little effect on our conclusions. We therefore set the issue of zero selections to the side for the rest of this thesis.

4.8 Chapter summary

Investigations outlined in this chapter first looked at the error rate and proportion of errors which were non-contextual on the first and second onsets individually. It was shown that the model tends to generate very high proportions of non-contextual errors on the second onset.

A statistical and graphical methodology for investigating the effect of manipulating parameter settings on the behaviour of the model was then presented, and this methodology is used throughout this thesis. The effects of parameter manipulations on error rate and the proportion of errors which were non-contextual on the first and second onset were outlined. Particularly significant findings included a demonstration that a higher number of steps before selection leads to a higher error rate, a result not in line with Dell's (1986) original claims, and results highlighting that overly high error rate and proportions of non-contextual errors can be caused both by connection strengths which are too low (c.f. Dell, Schwartz, et al., 1997) and connection strengths which are too high (c.f. Goldrick, 2006; Shrager et al., 1987). Crucially however, understanding the effect of parameter manipulations on basic behaviour of the model will aid the interpretation of their effects on more complex behaviour.

Limits on how high error rates and proportions of non-contextual errors can be for a specific model to be accepted as a model of human behaviour were then determined from corpus and experimental data. Analyses showed how many and which of the specific models tested met these criteria. Finally, it was demonstrated that over 99% of the specific models tested never abort productions due to zero selections.

The following chapters will use the parameter investigation methodology and the improved understanding of effects of parameters, and the classification of models as exhibiting or not exhibiting error generation behaviour in line with that demonstrated by normal human speakers, to aid evaluation of the model's ability to capture more complex evidence.

Chapter 5

Anticipations, perseverations and exchanges

5.1 Introduction

In the previous chapter, we examined the effects of manipulating parameters in the spreading activation model on basic behaviour, such as the error rate of a specific model, and the proportion of errors generated which are non-contextual. In this chapter, we focus solely on the contextual errors which the implementation can produce: anticipations (*"big fun"* → *"fig fun"*), perseverations (*"big fun"* → *"big bun"*), and exchanges (*"big fun"* → *"fig bun"*).

The ability of Dell's (1986) original model to replicate the relative proportions of anticipations, perseverations and exchanges reported in Nooteboom's (1969) corpus analysis, such that there were more anticipations than perseverations, and more perseverations than exchanges, forms an important part of the support for this model. However, there were some problems with Dell's (1986) comparison of model behaviour to empirical evidence. Firstly, as highlighted in chapter 2, other corpus analyses suggest different patterns of relative proportions of anticipations, perseverations and exchanges. Secondly, it appears likely that the high proportion of anticipations in Nooteboom's (1969) corpus was at least partly driven by the classification of all incomplete errors (*"big fun"* → *"fig. . . "*) as anticipations, when other authors have argued that these errors may in fact represent incomplete exchanges (e.g. Shattuck-Hufnagel, 1979). Thirdly, when investigating model behaviour, Dell (1986) did not impose any limits on how many errors the model could generate overall. As argued in chapter 4, there is an upper limit on the frequency with which normal speakers make errors, and good models of word production should observe this limit. Finally, it would be useful to clarify what effect the choice of spreading activation parameters is having on anticipation, perseveration and exchange generation. To what extent is the model's ability to replicate this evidence

due to the general architecture of the model, and to what extent is it reliant on a fortuitous choice of parameters? Importantly, given our planned multiple parameter setting analysis of the implementation's ability to account for the instrumental evidence summarised in chapter 2, we wish to uncover the parameter settings at which anticipation, perseveration and exchange generation operates appropriately in the spreading activation model, so that we can understand to what extent specific models which are able or unable to account for the new instrumental evidence can also explain this classic movement error evidence.

In this chapter, we begin by revisiting the corpus results reported in the literature to re-evaluate what sort of anticipation, perseveration and exchange proportion benchmarks we should be comparing the implementation's behaviour to. We then compare Dell's (1986) original results to the newly determined benchmarks, taking into account the limits on error rate and non-contextuality of errors which were established in the previous chapter. A new comparison of model behaviour and empirical results is then presented. In this new comparison, we investigate the effect of manipulating the spreading activation parameter settings on anticipation, perseveration and exchange generation and examine which sets of parameter settings allow the model to account for the empirical evidence as defined by the new benchmarks whilst observing the limits on error rate and non-contextuality of errors. We also explore the links between the variation in the implementation's behaviour caused by manipulations of the spreading activation parameters, and the empirical and theoretical investigations of the relationship between overall error rate and tendencies towards anticipation or perseveration error generation, as reported on by Dell, Burger, and Svec (1997).

5.2 Re-evaluating the behavioural evidence

In this section, we establish some new empirical benchmarks for relative proportions of anticipations, perseverations and exchanges, calculated from multiple corpora and based upon a careful consideration of the classification of incomplete errors. We then compare Dell's (1986) original results to these new benchmarks, taking into account the limits on error rate and non-contextuality of errors which were established in the previous chapter.

5.2.1 Establishing new benchmarks for relative proportions of anticipations, perseverations and exchanges

To establish new benchmarks on the relative proportions of anticipations, perseverations and exchanges generated by normal speakers, we first identify which speech error corpus reports provide the information we need to carry out a new analysis. Approaches towards classification of incomplete errors are then considered. Next, the results of two new analyses of the selected corpus reports are presented, and finally, we explain why experimental data has not been taken into account in these new analyses.

Identifying speech error corpus reports suitable for analysis

Dell (1986) only directly compared the proportions of anticipations, perseverations and exchanges his model generated to the proportions observed in one speech error corpus (Nooteboom, 1969). However, partially as a result of the two decades which have passed since Dell's (1986) paper was first published, there are now a number of other corpus reports available which provide analogous figures (Dell & Reich, 1981; del Viso et al., 1991; Garnham et al., 1981; Nooteboom, 1969, 2005b; Pérez et al., 2007; Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1989; Vousden et al., 2000).

Our comparison of Nooteboom (1969) and Shattuck-Hufnagel's (1979) results in chapter 2 showed that it is vital for us to understand the contribution of incomplete errors to reported relative proportions of anticipations, perseverations and exchanges. In our calculations of the new benchmarks, we therefore only referred to speech error corpora where the number of incomplete errors was explicitly specified. We found four corpus reports which provided this information (del Viso et al., 1991; Nooteboom, 2005b; Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1989). Nooteboom's (1969) corpus is not included, as no separate count of incomplete errors is provided. Table 5.1 describes the properties of these corpora, including the number of errors available for our analysis. Using the same approach as we did in chapter 4 when determining proportions of non-contextual errors in corpora, we selected the figures reported for the subset of errors closest to the onset consonant within-clause substitutions which our simulations will generate.

Table 5.2 shows the raw proportions of anticipations, perseverations, exchanges and incomplete errors in these corpora. This table emphasises how large a part of the corpora incomplete errors form, with proportions ranging from just under a quarter of all movement errors (del Viso et al., 1991) to nearly half the errors

recorded (MIT-CU corpus, Shattuck-Hufnagel & Klatt, 1979). Their classification as either anticipations or exchanges will therefore have a strong effect on the relative proportions of these two error classes.

Even before classification of incomplete errors however, it is clear that there is a lot of variation in the proportions of anticipations, perseverations and exchanges reported across the different corpora. One possible explanation of this variation could be that the sample size in these corpora is too small. However, we note that the number of errors considered in each of these analyses, ranging from 405 errors in del Viso et al.'s (1991) report to 1455 errors in Stemberger's (1989) report, is nearly as big as if not substantially bigger than the 535 errors considered in the analysis of Nooteboom's (1969) corpus used by Dell (1986), although of course the possibility remains that this variation would be reduced if more errors were collected. A more worrying explanation of the variation is that it is due to differing collector biases, such that some collectors are more sensitive to exchanges, and other to anticipations, and so on. However, as there would also be problems in considering experimental data as we outline later on, this is the best data currently available to us. More importantly, by pooling information from these differing corpora together and explicitly considering the influence of incomplete errors, this is arguably better data than the single report used in Dell's (1986) original important comparison, and it is worth clarifying how well the spreading activation model matches up to it.

The classification of incomplete errors

As noted in chapter 2, some previous analyses of speech error corpus data have chosen to classify all incomplete errors as anticipations (e.g. Nooteboom, 1969) or as exchanges (e.g. Shattuck-Hufnagel, 1979). The evidence for both of these extreme positions is limited however. The sole argument for classifying all incomplete errors as exchanges is provided by Shattuck-Hufnagel (1979), on the basis of data collected by Shattuck-Hufnagel and Klatt (as cited in Shattuck-Hufnagel, 1979), which suggests that the target and error phonemes involved in anticipations and perseverations are nearly always very similar, whereas this constraint is weaker for exchanges, and also incomplete errors. Classification of all incomplete errors as anticipations seems most likely to have resulted from the observation that an incomplete error involves all the error productions required for an anticipation, but not for an exchange. However, it can be argued that the correct production required for an anticipation also did not occur (e.g Cutler, 1981; Dell, 1986; Dell & Reich,

Table 5.1: Properties of speech error corpora used to determine boundaries for relative proportions of anticipation, perseveration and exchange errors

Name	Source paper	Collectors	Language	Number of errors in corpus	Errors considered for analysis	Number of errors in analysis[a]
Del Viso et al.	del Viso et al. (1991)	Two highly theoretically informed trained observers [b]	Spanish	3,611	Contextual phoneme substitution errors	405
Utrecht	Nooteboom (2005b)	Cohen, Nooteboom and others	Dutch	2,500	Contextual phonological substitution errors	1,153
MIT-CU	Shattuck-Hufnagel and Klatt (1979)	Shattuck-Hufnagel and Garrett	English	6,000	Contextual consonant substitution errors	1,134
Stemberger	Stemberger (1989)	Stemberger	English	4,000	Contextual within-clause phonological substitution errors	1,455

[a] Exchange errors are counted as one error.
[b] According to Pérez et al. (2007)

Table 5.2: Proportions of anticipations, perseverations, exchanges and incomplete errors in the speech error corpora used to determine new benchmarks

Name	Anticipations	Perseverations	Exchanges	Incomplete errors
Del Viso et al.	10.4%	46.9%	20.7%	22.0%
Utrecht	20.6%	22.2%	18.8%	38.3%
MIT-CU	10.0%	18.6%	24.3%	47.1%
Stemberger	17.3%	36.5%	5.0%	41.2%

1981; Fromkin, 1971; Garrett, 1975; Nooteboom, 1980, 2005b; Shattuck-Hufnagel, 1979; Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1989).

If we take the more moderate position that some incomplete errors represent incomplete anticipations, and some represent incomplete exchanges, how do we determine how many errors to allocate to each category? Two proposals have been put forward in the literature to address this problem. Nooteboom (2005b) hypothesises that speakers are equally likely to correct themselves after a perseveration as they are to correct themselves after an anticipation. On this premise, Nooteboom (2005b) determines the ratio of uncorrected perseverations to corrected perseverations, and uses this ratio in combination with the number of uncorrected anticipations to estimate the number of corrected anticipations; i.e., how many incomplete errors were actually incomplete anticipations. The remaining incomplete errors are deemed to have been exchanges. However, this methodology can only be applied to data where the numbers of corrected and uncorrected perseverations are given. Unfortunately, of the corpus reports we found in the literature, only Nooteboom (2005b) provides this information.

Another proposal is offered by Stemberger (1989) which requires only the number of incomplete errors to be explicitly provided in addition to the numbers of complete anticipations, perseverations and exchanges observed, as in the corpus reports summarised in tables 5.1 and 5.2. Stemberger (1989) suggests that the ratio of complete anticipations to complete exchanges is representative of the ratio of all anticipations to all exchanges. As such, incomplete errors should be split between the anticipation and exchange categories in this ratio. It should be noted, however, that this methodology is based on the assumption that you are no more likely to stop after an error where an upcoming word would also have been erroneous, than you are to stop after an error where the upcoming words would have been correct. If this assumption is incorrect such that speakers are more likely to stop when an error in the upcoming words is about to occur, then these calculations may still underestimate the underlying proportion of exchanges, whilst generously allocating

errors to the anticipation category. However, such an underestimate should be less extreme than that given by classifying all incomplete errors as anticipations. This approach is in fact also used by Dell (1986) when assessing his own experimental evidence.

A final proposal would be to ignore the incomplete errors completely, given that both Dell's (1986) model and our model in its wake do not have implemented error detection and editing systems, and therefore cannot generate incomplete errors. This approach would artificially boost the proportion of perseverations however, as these errors cannot be incomplete. On this basis, we reject this proposal for the current analyses.

We therefore consider two analyses of the four selected corpus reports (del Viso et al., 1991; Nooteboom, 2005b; Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1989). Our primary analysis follows the approach suggested by Stemberger (1989) and proportionally allocates incomplete errors across the anticipation and exchange categories. Without detailed information on corrections of perseverations, we believe that these calculations offer the most reasonable estimates of the underlying distribution of anticipations and exchanges in incomplete errors. Our secondary analysis categorises all incomplete errors as anticipations. This is the analysis used in Dell's (1986) original comparison of model behaviour and empirical results. By considering proportions determined in this second analysis we aim to maximise our understanding of the extent to which classification of incomplete errors is affecting the ability of the model to fit the human evidence. The results of these two analyses are presented below.

Two new analyses of relative proportions of anticipations, perseverations and exchanges in four corpora

In the first and principal analysis, the *proportional-incompletes* analysis, the ratio of complete anticipations to complete exchanges was calculated. Incomplete errors were then proportionally allocated to the anticipation and exchange categories (Stemberger, 1989) according to this ratio. For example, if there were three complete anticipations for every complete exchange, three incomplete errors were allocated to the anticipation category for every incomplete error allocated to the exchange category. The patterns of anticipations, perseverations, and exchanges uncovered by this analysis are shown in figure 5.1 and table 5.4.

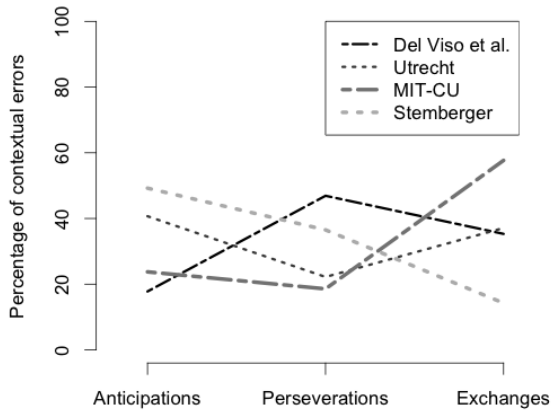


Figure 5.1: Analysis of corpora following the proportional-incompletes approach. See also table 5.3.

Table 5.3: Analysis of corpora following the proportional-incompletes approach. See also figure 5.1.

Name	Anticipations	Perseverations	Exchanges
Del Viso et al.	17.8%	46.9%	35.3%
Utrecht	40.7%	22.2%	37.1%
MIT-CU	23.8%	18.6%	57.7%
Stemberger	49.2%	36.5%	14.3%

Table 5.4: Analysis of corpora following the incompletes-as-anticipations approach. See also figure 5.2.

Name	Anticipations	Perseverations	Exchanges
Del Viso et al.	32.4%	46.9%	20.7%
Utrecht	59.0%	22.2%	18.8%
MIT-CU	57.1%	18.6%	24.3%
Stemberger	58.5%	36.5%	5.0%

Under the proportional-incompletes analysis, only the Stemberger (1989) data fits the numerical pattern originally reported by Nooteboom (1969) and modelled by Dell (1986), where there are more anticipations than perseverations, and more perseverations than exchanges. In the Utrecht corpus (Nooteboom, 2005b), anticipations are still the most common error, but there are more exchanges than perseverations. Exchanges are even more frequent in the MIT-CU corpus (Shattuck-Hufnagel & Klatt, 1979), with anticipations trailing behind, followed by perseverations. The del Viso et al. (1991) data gives rise to yet another pattern, in which perseverations have a higher reported occurrence than exchanges, which in turn are observed more often than anticipations.

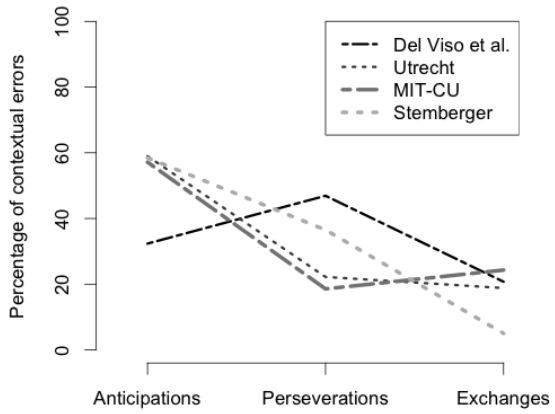
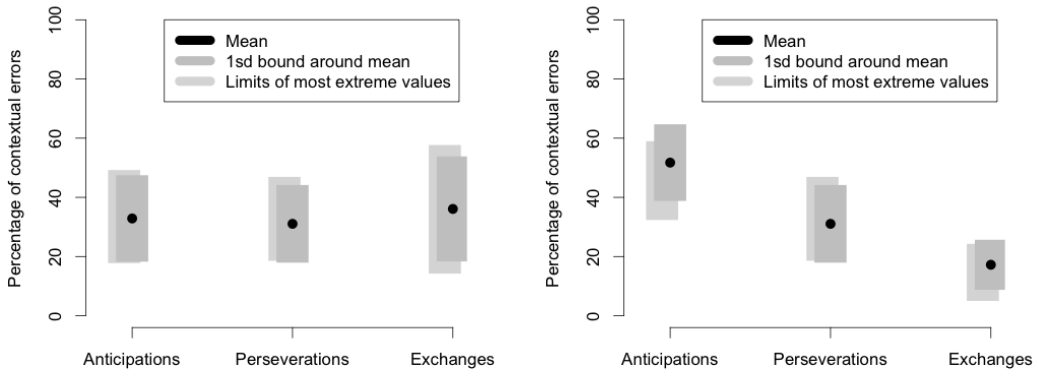


Figure 5.2: Analysis of corpora following the incompletes-as-anticipations approach. See also table 5.4.

The results of the secondary analysis, in which all incomplete errors are classified as anticipations (the *incompletes-as-anticipations* analysis) are shown in figure 5.2 and table 5.4. As in the proportional-incompletes analysis, the Stemberger (1989) data still shows the pattern reported by Nooteboom (1969), in which anticipations are more common than perseverations, and exchanges are least frequent of all. The reallocation of incomplete errors from the exchange category to the anticipation category means that the Utrecht data (Nooteboom, 2005b) now also fits this shape, although there are only marginally more perseverations than exchanges. In the MIT-CU corpus (Shattuck-Hufnagel & Klatt, 1979), the accumulation of incomplete errors in the anticipation category means that anticipations now appear more frequent than the previously most common error type, exchanges, which are still slightly more numerous than perseverations. Finally, perseverations are still the most commonly observed error in the del Viso et al. (1991) corpus, but with the extra incomplete errors, anticipations now have a slightly higher count than exchanges.

These results show that the claim that anticipations occur more often the perseverations, which in turn are claimed to occur more often than exchanges, is not an appropriate generalisation of the speech error patterns in the four corpora, regardless of whether incomplete errors are proportionally allocated to the anticipation and exchange category or are all counted as anticipations. In fact, no ordering of the frequency of occurrence of the three error types would generalise across all the data analysed. To attempt to usefully characterise the data so that we can assess



(a) (b)

Figure 5.3: Bounds on movement errors, calculated from corpora using the (a) proportional-incompletes and (b) incompletes-as-anticipations approaches. As explained in the text, the final upper and lower bound for each error types is considered to be the most extreme value of the standard deviation calculation or the raw data. See also tables 5.5 and 5.6.

whether simulations capture the broad patterns observed, we instead determined liberal bounds on the proportions of anticipations, perseverations and exchanges generated. The upper bound was one standard deviation above the mean, or the highest value reported from a corpus, whichever was greater, and similarly, the lower bound was one standard deviation below the mean, or the lowest value reported from a corpus, whichever was less.

The calculated bounds are depicted in figure 5.3 and described in tables 5.5 and 5.6. These results show that to meet the bounds determined from the proportional-incompletes analysis, none of the proportions for any of the error types should be extremely high or extremely low. For the incompletes-as-anticipations analysis, anticipation proportions should be high, perseverations middling, and exchange proportions low - yet not non-existent.

We note that neither of our analyses address the problem of classifying errors ambiguous between anticipations and perseverations, such as *"Well, long white hairs"* → *"Well, wong white hairs"* (Stemberger, 1989), where the /w/ in *wong* could have perseverated from *well* or been anticipated from *white*. Of the four corpora analysed here, only Stemberger (1989) explicitly reports the number of these ambiguous errors. For his data, we took the same approach as with the incomplete errors, and allocated the ambiguous errors to the anticipation and perseveration

Table 5.5: Bounds on anticipations, perseverations and exchanges using the proportional-incompletes analysis. Emboldened figures are the final bounds used, which are also summarised in italics at the bottom of the table. See also figure 5.3(a).

	Anticipations	Perseverations	Exchanges
Mean	*32.9%*	*31.1%*	*36.1%*
1 s.d. below mean	18.3%	**17.9%**	18.3%
1 s.d. above mean	47.5%	44.2%	53.8%
Lowest reported value	**17.8%**	18.6%	**14.3%**
Highest reported value	**49.2%**	**46.9%**	**57.7%**
Final most liberal lower bound	*17.8%*	*17.9%*	*14.3%*
Final most liberal upper bound	*49.2%*	*46.9%*	*57.7%*

Table 5.6: Bounds on anticipations, perseverations and exchanges using the incompletes-as-anticipations analysis. Emboldened figures are the final bounds used, which are also summarised in italics at the bottom of the table. See also figure 5.3(b).

	Anticipations	Perseverations	Exchanges
Mean	*51.7%*	*31.1%*	*17.2%*
1 s.d. below mean	38.8%	**17.9%**	8.8%
1 s.d. above mean	**64.7%**	44.2%	**25.7%**
Lowest reported value	**32.4%**	18.6%	**5.0%**
Highest reported value	59.0%	**46.9%**	24.3%
Final most liberal lower bound	*32.4%*	*17.9%*	*5.0%*
Final most liberal upper bound	*64.7%*	*46.9%*	*25.7%*

categories in the same ratio as complete anticipations and perseverations occur. This calculation was done prior to allocating the incomplete errors, and as such, some ambiguous errors counted towards the anticipation tally when determining the ratio of anticipations of exchanges. The effect of these ambiguous errors on the anticipation and perseveration counts in other corpus data remains unaccounted for, however.

Why not use experimental results as well as corpus data?

The analysis of human anticipation, perseveration and exchange generation presented here considers multiple speech error corpora instead of just one. However, a wealth of experimental speech error evidence also exists in the literature, which we have not taken into account. This decision stemmed from concerns about the effect of laboratory error elicitation techniques on the data we are interested in.

As humans make speech errors so infrequently in natural settings, experimenters are forced to rely on error boosting paradigms to make laboratory data collection practical. These paradigms include the tongue twister task, in which participants are asked to produce sequences of words such as *palm neck name pack* (for a review, see Wilshire, 1999). Ordering confusion between the repeated onset consonants, and sometimes similarity of the onset phonemes, is intended to induce contextual errors on the onsets of the words. In another error elicitation paradigm, the Word Order Competition (WOC) task (Baars & Motley, 1976), participants are shown pairs of words followed by an arrow pointing left or right, and are instructed to produce the presented pair in the direction indicated by an arrow. A leftward pointing arrow is used for all target pairs and means that the participant must reverse the order of the words in the pair (for example, the presented target *rain gate* should be produced *"gate rain"*). Confusion between the reversed word order which the participant is instructed to produce and the original presented order of the words is intended to cause the participant to exchange the onset consonants of the target pair, to produce *"rate gain"*. A third common experimental technique is the use of the SLIP task (e.g. Baars et al., 1975), where participants are sequentially presented with a series of priming word pairs for silent reading, such as *give rust, gale raise*, and are then prompted to produce a target pair such as *rain gate*, in which the onset consonants are reversed in comparison to the priming pairs. The paradigm thereby again aims to induce the participant to mistakenly exchange the consonants of the target pair, in this example resulting in *"gain rate"*.

However, identifying errors produced in these paradigms as anticipation, perseverations or exchanges can be problematic. The tongue twister task is particularly susceptible to the ambiguous anticipation/perseveration classification issue, as competing phonemes normally occur multiple times in an utterance. For example, if the tongue twister *palm neck name pack* is misproduced as *"palm peck name pack"*, it is not possible to determine whether the /p/ in *peck* is a perseveration from *palm* or an anticipation from *pack*. In the WOC task, it is often unclear what the participant really intended to say when the error occurred. For example, according to the task instructions, a target pair *rain gate* should be reversed to *"gate rain"*, but a participant may instead produce the error *"gain gate"*. If we assume that the participant really intended to produce *"gate rain"*, then *"gain gate"* would be classified as a rime exchange, and an onset perseveration. However, it could also be argued that the participant actually tried to produce the presented pair *"rain gate"*, and that the error *"gain gate"* is a simple onset anticipation. It would be difficult to demonstrate that either of these two conflicting classifications was the correct

choice. Finally, if the preceding priming pairs in the SLIP task are presumed to be the cause of the increased error rate on target pairs, it can be argued that all the resulting errors are to some extent perseverations. On the basis of these concerns, our analyses remain restricted to the patterns reported in speech error corpora.

5.2.2 Comparing Dell's (1986) original simulation results to the new speech error corpus benchmarks

To relate the benchmarks determined in the previous section to Dell's (1986) original investigations, we first revisit the comparison that Dell (1986) made between his simulation results and the speech error corpus results reported by Nooteboom (1969), and then compare Dell's (1986) simulations to the new benchmarks.

Dell's (1986) original results and Nooteboom's (1969) corpus data

Dell (1986) investigated the proportions of anticipations, perseverations and exchanges generated by his model by simulating the production of 120 word pairs. These were randomly selected from the model's vocabulary of 50 two-syllable common English words, none of which were more than eight letters long. The model was tested with three timesteps before phoneme selection, four timesteps before selection, and eight timesteps before selection. Dell (1986) then compared the proportions of anticipations, perseverations and exchanges generated by his simulations to the proportions of these errors found in Nooteboom's (1969) adult speech error corpus.

Table 5.7 shows the results from Dell's (1986) simulations, and summarises the corpus data reported by Nooteboom (1969). As analyses of our simulations will focus on onset phoneme errors, the data reported for both the simulations and the corpus data relates to phoneme substitution errors only. This data does not differ greatly or critically to the data for errors of all unit sizes however.

Dell (1986) reported that a number of errors generated by the model were ambiguous between anticipations and perseverations. For example, the error *"infant urchin"* → *"infin urchin"* could be classified as an anticipation, where the *in* from *urchin* had been anticipated; or a perseveration, where the *in* from *infant* had been perseverated. For comparison of the model's behaviour to Nooteboom's (1969), Dell (1986) split these errors equally between the anticipation and perseveration categories.

Table 5.7: Dell's (1986) simulation results and Nooteboom's (1969) corpus data

		Anticipations	Perseverations	Exchanges
	Timesteps before selection			
Simulation	3	34.3%	57.8%	7.8%
	4	62.3%	34.0%	3.8%
	8	90.9%	9.1%	0.0%
Nooteboom (1969)		76.1%	17.8%	6.2%

These results show that when there are four or eight timesteps before selection, the model generates more anticipations than perseverations, and more perseverations than exchanges, as is the pattern found in Nooteboom's (1969) corpus data. In contrast, as Dell (1986) highlights, when only three timesteps pass before selection, the model behaviour pattern does not match the results found by Nooteboom (1969), such that there are more perseverations than anticipations. However, exchange errors are still the least frequent of the three error types.

Dell (1986) observes that, as the most complex of the three error types, "exchanges are necessarily less common than anticipations or perseverations in the model" (p. 300; see also p. 292). However, it should be noted that with eight timesteps before selection, there are actually no exchange errors at all. With four timesteps before selection, 3.8% of phoneme substitution errors are exchange errors, barely more than half of the 6.7% observed by Nooteboom (1969). Exchange error proportions are notably healthier with three timesteps before selection, where 7.8% of phoneme substitutions are exchanges. Yet at this setting, the overall pattern does not reflect Nooteboom's (1969) data, such that the simulated proportion of perseverations is more than three times the proportion seen in the corpus report, and is vastly greater than the anticipation proportion, which is about half the anticipation proportion observed by Nooteboom (1969).

Dell's (1986) original results and our newly determined benchmarks

Table 5.8 compares the behaviour of Dell's (1986) simulations to the liberal upper and lower bounds we determined in section 5.2.1 on the relative proportions of anticipation, perseveration and exchange errors generated. These results show that no timestep setting allowed the model to meet the bounds for all error types regardless of the approach to incomplete error classification. An acceptable proportion of anticipations was generated at three timesteps, the setting which in Dell's (1986) comparison to Nooteboom's (1969) was least successful, but at four timesteps the

Table 5.8: Comparison of the behaviour of Dell's (1986) model to new empirical bounds

Timesteps before selection	Analysis	Anticipations	Perseverations	Exchanges
3	PI	ok	too high	too low
	IAA	ok	too high	ok
4	PI	too high	ok	too low
	IAA	ok	ok	too low
8	PI	too high	too low	too low
	IAA	too high	too low	too low

Key: PI = proportional-incompletes, IAA = incompletes-as-anticipations

proportion was too high under our primary proportional-incompletes analysis, and at eight timesteps there were too many anticipations under all analyses. The perseveration proportion was good at four timesteps, but too high with fewer timesteps and too low with more. Yet the most pervasive problem was exchange error generation. Under nearly all analyses and timestep settings, too few exchange errors were generated to meet our liberal bounds. The one exception to this result was again for the three timestep setting. This simulation just scraped the bounds determined from the secondary incompletes-as-anticipations corpus analysis, in which the proportion of anticipations is increased at the expense of exchanges.

This problem is worsened when we take into account the 5.75% error rate upper bound, as determined in section 4.5, and compared to the error rates estimated for Dell's (1986) simulations at each of the timestep settings in section 4.4.3, such that with three timesteps before phoneme selection, there were approximately 16.9% errors; with four timesteps, 5% errors; and with eight timesteps, 4.4% errors.

These results suggest that the model may have remained under our very liberal error rate upper limit of 5.75% when there are four or eight timesteps before phoneme selection, but it looks unlikely that it was successful when there were only three timesteps before phoneme selection, which is the only setting at which the simulation produced enough exchanges to fall within the bounds of our most conservative estimate of exchange error proportions. This is particularly worrying for the model because as error rate rises, the probability of double errors (i.e., exchange errors) occurring by chance rises and the probability of single errors accompanied by a correct production (i.e., anticipations and perseverations) occurring by chance falls, such that the proportion of exchange errors generated will naturally increase. There is therefore some suggestion that the model is relying on inappropriately high error rates to generate sufficiently high proportions of exchange errors. Non-contextuality

of errors is, on the other hand, not a problem in Dell's (1986) results, as Dell (1986) reports that his model makes no non-contextual errors at all (although this is somewhat surprising; see earlier discussion in section 4.4.3).

5.2.3 Behavioural evidence re-evaluation summary

In this section, we identified four speech error corpus reports in the literature where the number of incomplete errors is explicitly reported (del Viso et al., 1991; Nooteboom, 2005b; Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1989), rather than the number of incomplete errors being integrated into the proportions of anticipations, perseverations and exchanges in a non-transparent fashion, as in the corpus to which the behaviour of Dell's (1986) model was originally compared (Nooteboom, 1969). We noted that in all of these corpus reports, incomplete errors form between a quarter and half of the errors recorded. Their classification as either anticipations or exchanges therefore has a strong influence on the relative proportions of error types. However, the proportions of anticipations, perseverations and exchanges do differ quite considerably between these reports even before the classification of incomplete errors is considered.

We presented two analyses of the data in these corpus reports. In the primary analysis, the *proportional-incompletes* analysis, some incomplete errors are classified as anticipations, and others are classified as exchanges, following a method proposed by Stemberger (1989). In the secondary analysis, the *incompletes-as-anticipations* analysis, all incomplete errors are categorised as anticipations, as in Nooteboom's (1969) original data. The bounds determined from the proportional-incompletes analysis specify that none of the proportions for any of the error types should be extremely high or extremely low. The bounds determined from the incompletes-as-anticipations analysis, suggest that anticipation proportions should be high and exchange proportions low, although not non-existent.

We then showed that Dell's (1986) original simulation results do not meet these newly determined bounds, regardless of how many timesteps pass before phoneme selection. In particular, the model appeared to struggle with exchange error generation, such that too few were generated. No timestep settings led the model to generate enough exchanges to meet the lower bound on exchange error proportions determined from the primary proportional-incompletes analysis. Only with three timesteps before phoneme selection could the model generate enough exchanges for the much lower bound determined from the secondary incompletes-as-anticipations

analysis. However, the overall error rate at this setting was very high at 16.9%, which is far above the upper limit on error rate determined in section 4.5 of 5.75%. As the proportion of exchanges generated by chance should rise as the error rate rises, this result raises the worrying suggestion that the model is relying on inappropriately high error rates to generate sufficiently high proportions of exchanges.

In next sections, we clarify the effect of manipulating the spreading activation parameter settings on the proportions of anticipations, perseverations and exchange errors generated, and investigate whether other parameter settings permit the model to be more successful at meeting these new bounds.

5.3 Simulation methodology

5.3.1 Model configuration, lexicon and task

The results reported in this chapter were derived from the same random word pair production simulations as the results reported in the previous chapter. The model configuration, lexicon and task are therefore all the same as in the previous chapter. To understand which parameter settings allow the model to account for this evidence, and then use this information to inform our later investigation of information flow between phonological and subphonemic processing stages and the model's ability to account for the instrumental evidence summarised in chapter 2, we would need to consider the behaviour of the two-stage model and test all connectivity settings which are tested in the later simulations. However, we began by focusing on evaluating the effect of varying the parameter settings within a one-stage phonological encoding model with output from the phoneme level, with feedback from phonemes to words, and feedback from features to phonemes, as was used by Dell (1986).

5.3.2 Model output classification

Productions on individual onsets are classified as correct productions, contextual errors or non-contextual errors as in the previous chapter. This simulation focused on the classification of word pair productions. Word pair productions can be classified as correct word pair productions, anticipations, perseveration, exchanges, or non-contextual word pair errors,and this classification is based on the classification of the onset productions. A correct word pair production is recorded if the first onset and the second onset are both correctly produced. An anticipation is recorded

if there is a contextual error at the first onset, followed by a correct production at the second onset. A perseveration is recorded if the first onset is correctly produced, but there is a contextual error at the second onset. An exchange is recorded if there is a contextual error at the first onset, followed by a contextual error at the second onset. Finally, if either the first or second onset production results in a non-contextual error, a non-contextual word pair error is recorded, regardless of the status of the other onset.

Note that as the model produces phrases of two syllables only, it is not possible for contextual errors to be ambiguous between anticipations and perseverations, as contextual errors on the first syllable are assumed to have a source in the second syllable, and vice versa.

5.4 Simulation results

In this section, we compare the anticipation, perseveration and exchange error generation behaviour of the model to the empirical bounds determined in section 5.2.1, whilst taking into account the limits on error rate and non-contextuality of errors outlined in section 4.5. Using the graphical and statistical approach described in section 4.4.1, we elucidate the effects of manipulating the spreading activation parameters on the implementation's anticipation, perseveration and exchange generation behaviour. Again, chi-squared tests on the likelihood ratio tests showed that all parameters always made highly significant contributions ($p < 0.0001$) to models of every measure considered in this chapter. We therefore take the same approach as in the previous chapter and do not comment further on the significance of the parameter contributions to the models, instead focusing on exploring the comparative size of effects as indicated by the calculated Wald's Z value.

We first consider the implementation's generation of anticipations and perseverations and the effect of the parameter manipulations on the proportions of these errors generated. The relationship between the proportion of errors which are anticipatory rather than perseveratory and the overall error rate is then described, and these results are compared to Dell, Burger, and Svec's (1997) empirical and theoretical findings. Finally, the implementation's ability to generate exchange errors is investigated, and the influence of the spreading activation parameter settings on these results is clarified.

Of the 5832 specific models tested, 1727 produced only correct pairs. For a further 13 specific models, all the word pair errors were deemed non-contextual according to the criteria set out in section 5.3, due to the errors on one or both of the onsets being non-contextual. This left 4092 specific models whose word pair error behaviour could be measured against our anticipation, perseveration and exchange constraints, 70.2% of all the specific models tested. Of these models, 1665 specific models (28.5% of the models tested) did not generate too many errors or too high a proportion of non-contextual errors for the limits on these measures established in section 4.5.

5.4.1 Anticipations and perseverations

We begin with the anticipation and perseveration proportions.

Overview of implementation behaviour

The behaviour of the model was first examined across all models which generated contextual errors for analysis. Figure 5.4 shows that the 4092 specific models analysed contained models which simultaneously generated appropriate amounts of both anticipations and perseverations, for both the primary proportional-incompletes and secondary incompletes-as-anticipations analyses. Specifically, the proportion of contextual word pair errors which were perseverations fell within the bounds for the perseveration constraint (common to both analyses) on 527 specific models, 9.0% of the models tested. Of these 527, 33 generated a proportion of anticipations which fell within the proportional-incompletes analysis bounds (0.6% of models tested), and 271 generated a suitable proportion of anticipations for the higher secondary incompletes-as-anticipations bounds (4.6% of models tested). However, the rest of the specific models which generated appropriate proportions of perseveration errors displayed too high a proportion of anticipation errors.

Similarly, the largest portion of the 4092 specific models being analysed, comprising 2097 specific models in total (36.0% of models tested) displayed both too high a proportion of anticipation errors and too low a proportion of perseverations for both analyses. However, another sizeable portion of the models, composed of 1468 specific models (25.2% of models tested) produced too high a proportion of perseveration errors. Most of these also generated too few anticipations, with 807 specific models (13.8% of all models tested) falling into this category according to the proportional-incompletes analysis, and 1106 specific models (19.0% of models

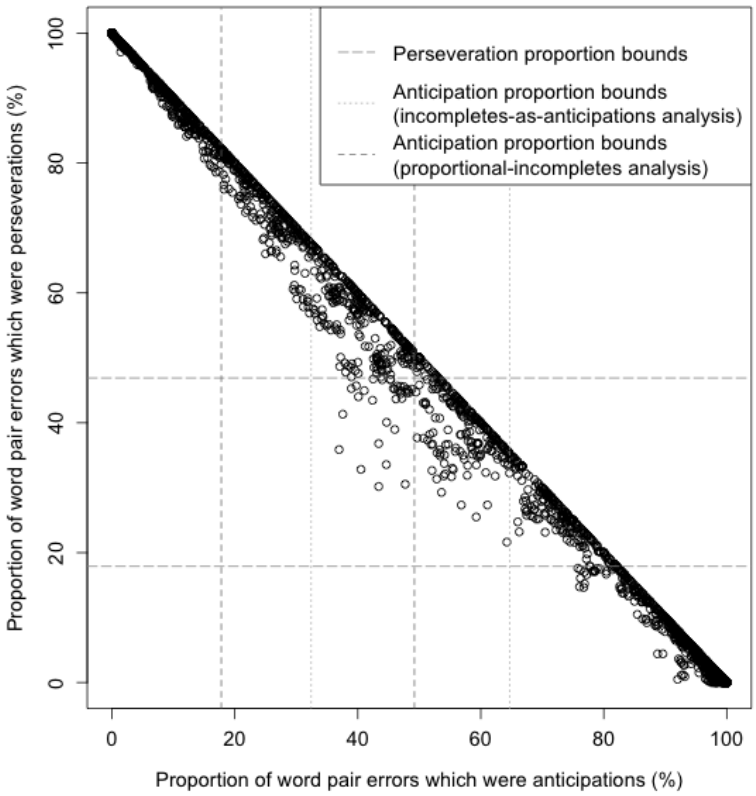


Figure 5.4: The proportion of contextual word pair errors which were anticipations plotted against the proportion of contextual word pair errors which were perseverations, for all specific models which generated at least one anticipation, perseveration or exchange

tested) generating too few anticipations for the incompletes-as-anticipations analysis.

The model had even less success when we considered only the 1665 specific models which both generate errors and pass the constraints on error rate and non-contextuality of errors. Figure 5.5 shows that within these specific models, no specific models simultaneously generated appropriate proportions of both anticipations and perseverations given the proportional-incompletes analysis bounds, and only a few specific models generated proportions of anticipations and perseverations which fell within the bounds specified by the secondary incompletes-as-anticipations analysis. Specifically, 70 of the 1665 specific models (1.2% of all models tested) generated an appropriate proportion of perseverations for the common perseveration bounds, but all of these 70 specific models generated relatively too many anticipations for

the proportional-incompletes analysis. For the incompletes-as-anticipations analysis, 15 specific models displayed a proportion of anticipation errors which fell within the specified bounds (0.3% of all models tested), but the proportion of contextual word pair errors which were anticipations was again too high for the remaining 55 specific models.

By far the most pervasive problem for the anticipation and perseveration constraints, affecting 1359 of the 1665 error generating specific models which passed the constraints on error rate and non-contextuality of errors (23.3% of all models tested), was too low a proportion of perseverations alongside too high a proportion of anticipations for both analyses.

The remaining 236 specific models displayed too high a proportion of perseverations (4.0% of all models tested), though clearly many specific models which generated too many perseverations had been ruled out by the constraints on error rate and non-contextuality of errors. For the proportional-incompletes analysis, 166 of these specific models generated relatively too few anticipations (2.8% of models tested), and for the secondary incompletes-as-anticipations analysis, 185 of these 236 specific models generated too low a proportion of anticipations (3.2% of all models tested).

To summarise, when analysing the 70.2% of all models tested which generated contextual errors for analysis, 0.6% of all models tested generated appropriate proportions of anticipations and perseverations to meet the proportional-incompletes bounds, and 4.6% met the incompletes-as-anticipations bounds.

The most prevalent problem was specific models generating too many anticipations and too few perseverations, affecting 36.0% of all models tested. However, 13.8% of models showed the reverse problem of too many perseverations and too few anticipations according to the proportional-incompletes analysis, a figure which rose to 19.0% of models under the incompletes-as-anticipations analysis.

Once models which did not meet the error rate and non-contextuality constraints were excluded however, leaving 28.5% of all models tested available for analysis, only 0.3% of all models tested met the secondary incompletes-as-anticipations analysis constraints on proportions of anticipations and perseverations generated, and no specific models met the proportional-incompletes analysis bounds.

Many of the models which passed the error rate and non-contextuality constraints generated too many anticipations and not enough perseverations, with 23.3% of all models tested meeting this description. Conversely, our results showed that a large

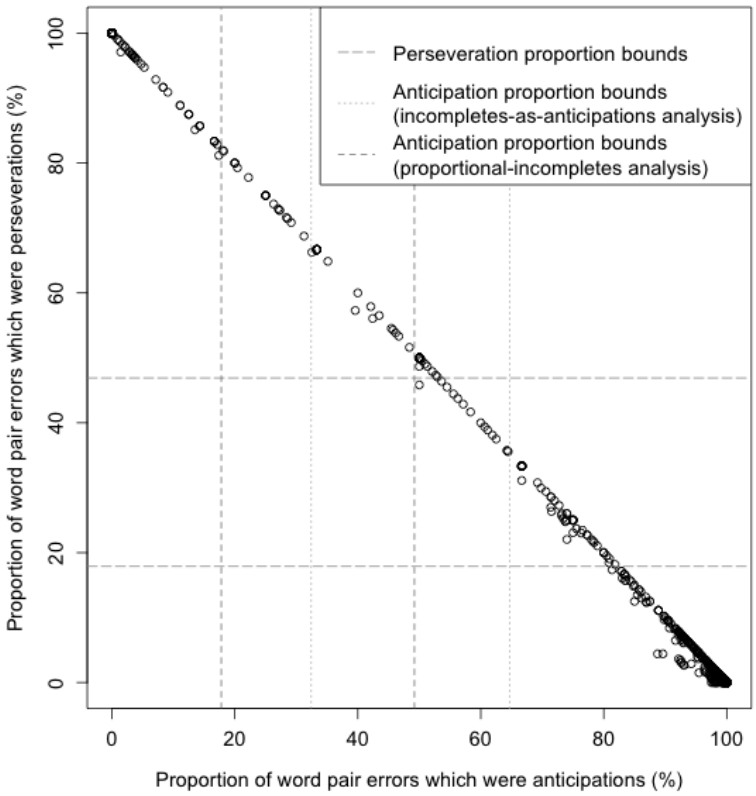


Figure 5.5: The proportion of contextual word pair errors which were anticipations plotted against the proportion of contextual word pair errors which were perseverations, for all specific models which passed both constraints on overall error rate and non-contextuality of errors

proportion of the models which generated too many perseverations and not enough anticipations were ruled out by the error rate and non-contextuality constraints, with only around 2.8% of all models tested both passing the error rate and non-contextuality constraints and showing this pattern according to the proportional-incompletes bounds, or 3.2% of all models tested when the incompletes-as-anticipations bounds were used.

These rather worrying results show that the implementation struggles greatly to meet even the secondary incompletes-as-anticipations bounds on anticipation and perseveration generation, and once the constraints on error rate and non-contextuality of errors are taken into account, cannot meet the proportional-incompletes bounds at all. However, in considering these error proportion results, we have not yet taken exchange error generation into account. For both analyses, the maximum accepted proportion of perseverations is 46.9%. Under the primary proportional-incompletes

analysis, the maximum accepted proportion of anticipations is 49.2%, so 3.9% of contextual word pair errors must be exchanges for both the anticipation and perseveration bounds to be met. For the secondary incompletes-as-anticipations analysis on the other hand, the higher maximum anticipation proportion of 64.7% means that the simulations can still meet the anticipation and perseveration bounds even without generating any exchanges. Our analysis of Dell's (1986) original results in section 5.2.2 showed that in Dell's (1986) original simulations, the model generated very few exchanges. Similar behaviour here would help explain why the model tends towards generating too many anticipations and not enough perseverations, or vice versa, and also why the model has particular difficulty meeting the bounds of the proportional-incompletes analysis. We will therefore return to this point in section 5.4.2 where the proportions of exchange errors generated by the model are analysed.

In this section, we focus on those anticipation and perseveration results that are largely independent of exchange error generation. In particular, we investigate why excluding the simulations which generate too many errors or too high a proportion of non-contextual errors rules out many of the simulations which generate too many perseverations and too few anticipations, but does not greatly affect the simulations which generate too many anticipations and not enough perseverations. The results reported in the previous chapter offer some insight into this result. In section 4.3, we observed that for the majority of simulations, a high proportion of second onset errors are non-contextual, whereas most first onset errors are contextual. We argued that this is presumably due to contextual errors on the first onset being directly primed, whereas contextual error generation on the second onset relies on activation of neighbours in the network, such that non-contextual representations are more likely to have also received some activation. Given this result, a bigger increase in overall error rate would therefore be expected to accompany an increase in contextual second onset errors than would be the case for contextual first onset errors. Similarly, when second onset errors make up a larger part of the errors generated, the overall proportion of non-contextual errors will rise, as second onset errors are more frequently non-contextual. Following this logic, specific models with more second onset contextual errors, and therefore perseverations, are more likely to be excluded by the error rate and error non-contextuality constraints.

In the previous chapter, we looked at the effects of manipulating the spreading activation parameter settings on first onset and second onset error rate. In this section, we extend this work by investigating the effect of parameter manipulations on the number of anticipations and perseverations generated. We also further consider the

Table 5.9: Results of logistic regression model analyses using parameter values to predict the percentage of word pair productions which were anticipations. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	+	41.9	1745	< .001	*
joltPrimeRatio	−	813.7	2350484	< .001	*
decay	+	176.6	31354	< .001	*
steps	+	744.6	614656	< .001	*
actiNoiseSD	+	786.9	696657	< .001	*
intrinNoiseSD	+	14.4	207	< .001	*

result that specific models that generate a high proportion of perseverations are often excluded by the constraints on error rate and non-contextuality of errors, by investigating to what extent this result can be linked back to Dell, Burger, and Svec's (1997) empirical and theoretical considerations of the relationship between anticipation and perseveration generation and overall error rate.

Effects of parameter manipulations on anticipation and perseveration generation

We focus here on the effect of parameter manipulations on anticipations and perseverations. We consider what percentage of all word pairs are produced as anticipations, or perseverations, rather than looking at the proportion of contextual word pair errors that are anticipations or perseverations, as proportion measurements are so strongly affected by the rate of occurrence of other word pair error types. Our understanding of the effects of parameter manipulations on individual error type occurrence rates can then be used to work out what error types are driving changes in proportions, as we demonstrate in the next section.

The logistic regression summarised in table 5.9 and the graphs in figure 5.6 show that the parameters which were demonstrated to lead to an increase in first onset errors in the previous chapter also lead to an increase in anticipation errors. Specifically, higher connectivity strengths, lower jolt to prime ratios, higher decay rates, higher numbers of timesteps before phoneme selection, and higher activation-based and intrinsic noise levels all increase the number of anticipation errors generated.

The explanations provided in section 4.4.2 for the directionality of these effects on first onset error generation similarly apply to anticipation generation. Higher connection strengths increase the extent to which the network is swamped with

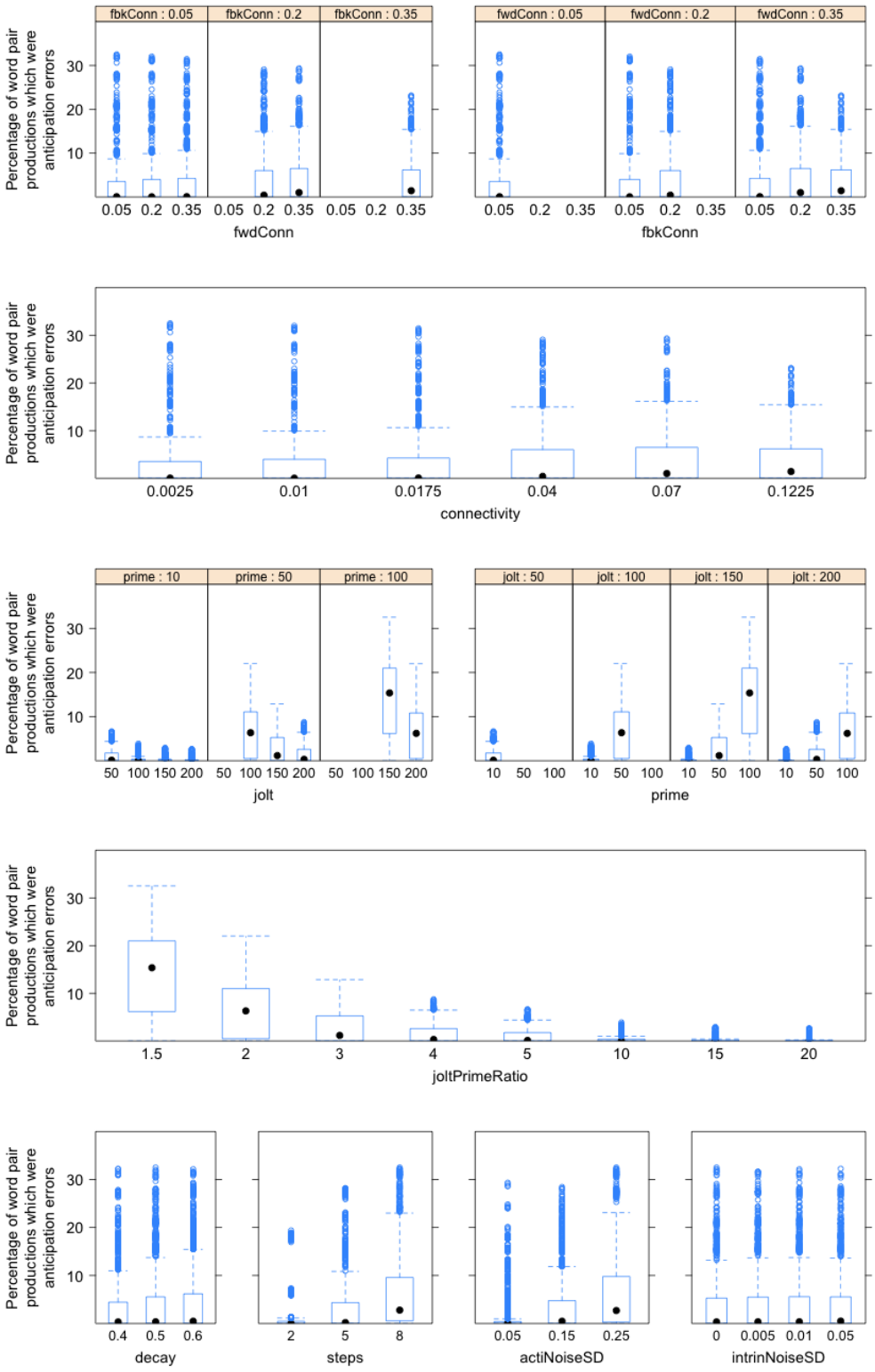


Figure 5.6: The effect of changing parameter settings on the percentage of word pair productions in each specific model which were anticipation errors, for all specific models.

activation, making errors more likely. Figure 5.6 shows that some low connection strength simulations demonstrate some of the higher rates of anticipation error occurrence however. This is due to a reduced influence of the jolt activation which cannot pass from the jolted word to the target phoneme as effectively when connection strengths are low. Low jolt to prime ratios mean that the prime activation of the upcoming second onset is comparatively higher, increasing the likelihood of an early selection. High decay rates lead to the jolt activation decaying more rapidly, so that errors are more likely. A higher number of timesteps before phoneme selection has a similar effect, providing more opportunity for the jolt activation to decay, and noise to affect the activation patterns in the network and cause errors to occur. Finally, higher noise levels also reduce the influence of the jolt activation and make errors more likely.

Looking at the importance of different parameters as depicted by the Wald's Z values in table 5.9, it is clear that the jolt to prime ratio has the biggest effect on the number of anticipations generated. This is because low jolt to prime ratios massively increase the number of first onset errors generated, but not the number of second onset errors. The level of activation-based noise also has a strong influence as higher levels of activation-based noise have a particularly strong effect on the primed second onset, making it more likely that this phoneme is anticipated. As for all measures in the previous chapter, increasing the number of steps before selection also has a very strong effect, such that more anticipations are generated when the number of steps before selection is higher. However, the effect of the timesteps setting, the connectivity strength and the level of intrinsic noise are all less important for anticipation generation than they are for first onset error generation because increasing these parameters also leads to a rise in second onset errors, such that the anticipation pattern of an error on the first onset followed by a correct second onset is not so strongly promoted. Finally, increasing the decay rate leads to a mild increase in the number of anticipations generated, but this effect is weak as it is on individual first onset error rates.

The regression model in table 5.10 and the graphs in figure 5.7 show that it is also on the whole true that the parameters which were demonstrated to lead to an increase in second onset errors in the previous chapter lead to an increase in perseveration errors, with the exception of the effect of jolt to prime ratio. Specifically, an increase in connectivity strength, a decrease in decay rate, an increase in the number of steps before phoneme selection, and increases in the levels of activation-based and intrinsic noise all lead to higher numbers of both second onset errors and

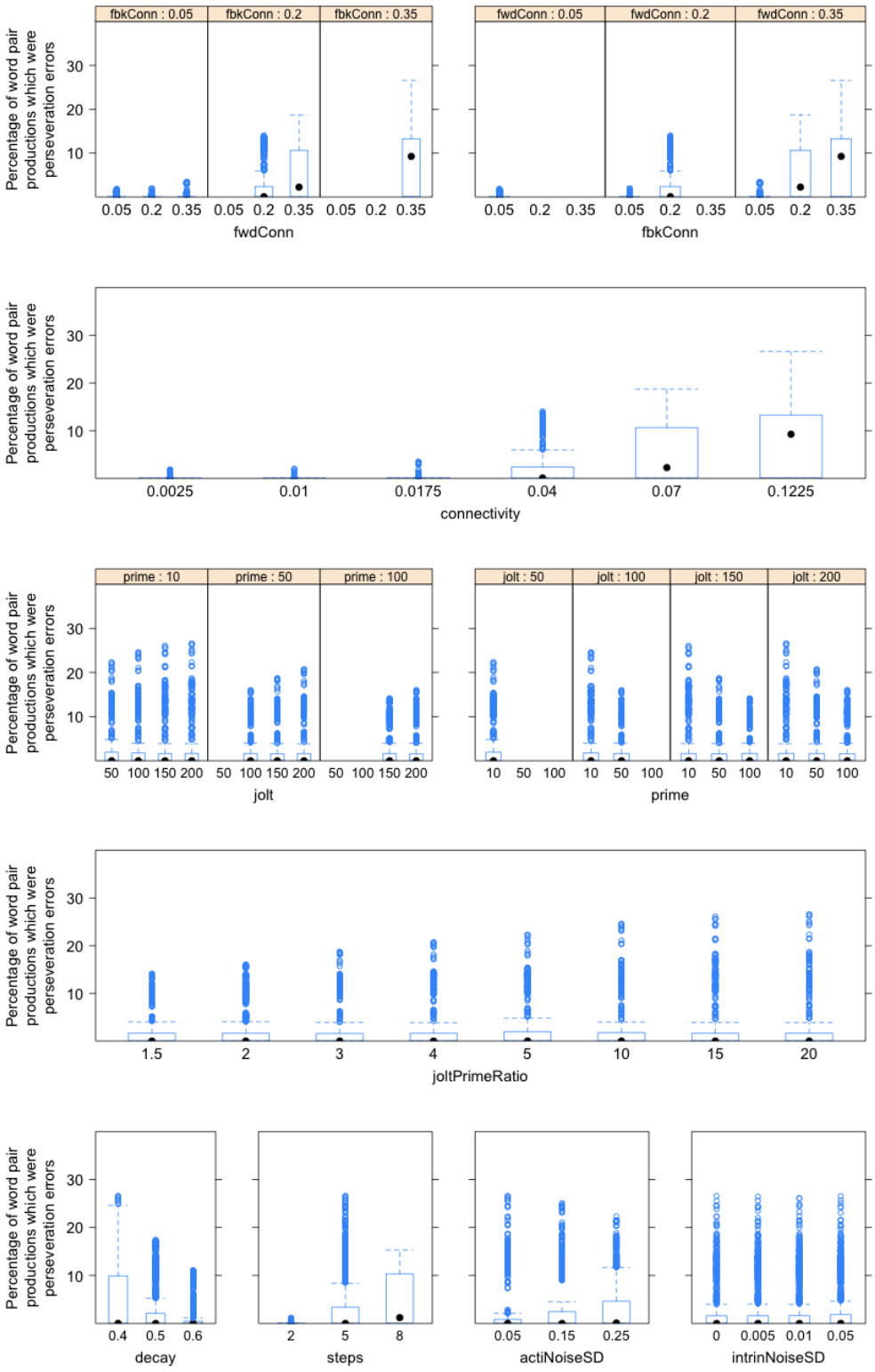


Figure 5.7: The effect of changing parameter settings on the percentage of word pair productions in each specific model which were perseveration errors, for all specific models.

Table 5.10: Results of logistic regression model analyses using parameter values to predict the percentage of word pair productions which were perseverations. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	+	1229.7	1777949	< .001	*
joltPrimeRatio	+	121.7	14467	< .001	*
decay	−	609.8	411490	< .001	*
steps	+	834.0	879437	< .001	*
actiNoiseSD	+	59.1	3491	< .001	*
intrinNoiseSD	+	12.5	156	< .001	*

perseverations. However, whilst a higher jolt to prime ratio causes less errors to be generated on the second onset, it leads to more perseverations occurring.

As for anticipations, higher connection strength generally increases the amount of activation in the network, reducing the influence of the jolt activation. However, for perseverations, higher connection strengths also aid the reactivation of previously produced first onsets. For example, for the target phrase *"big fun"*, higher connection strengths allow neighbours of *big*, such as *bill* and *bat* to become highly activated during first word production, and then also facilitate the passage of activation from these neighbours back to the onset phoneme /b/ during second word production, increasing the likelihood of the perseveration *"big bun"*. We note however that whilst some specific models with low connection strengths demonstrated high second onset error rates, there is little evidence of perseveration generation increasing at these settings. As low connection strengths would not support neighbour reactivation, it seems likely most of the errors generated by low connection strength specific models were non-contextual, an argument supported by the high proportions of non-contextual errors reported for these models.

Lower decay rates mean that activation on neighbours decays more slowly, thereby also increasing perseveration rate. As for anticipations, higher numbers of timesteps before phoneme selection cause the jolt activation to decay, so that the effect of noise on the activation levels of representations is stronger, and more errors occur. Note again that this result is contrary to Dell's (1986) result that more perseverations occur when fewer timesteps pass before phoneme selection, for the reasons outlined in detail in section 4.4.3. Finally, higher levels of activation-based and intrinsic

noise on representations reduce the influence of the jolt activation and increase the probability of an error.

However, a higher jolt to prime ratio leads to more perseverations occurring, despite the fact that higher jolt to prime ratios lead to small reductions in second onset error rate. In section 4.4.3, we argued that whilst higher primes at lower jolt to prime ratios cause more second onset errors overall, a higher jolt to prime ratio was more conducive to perseveration generation. Again, consider production of the phrase *"big fun"*. At a high jolt to prime ratio, it is more likely that the first onset will be produced correctly. Following a correct first onset production, a lower prime will mean that the activation of the second onset is lower than it would otherwise have been relative to the rest of the network (in particular, the neighbours of *big*, such as *bill* and *bat*). The activation passing back to the first onset /b/ from neighbours of *big* during production of the second word will therefore be comparatively stronger, making production of the perseveration *"big bun"* more likely.

The Wald's Z values reported in table 5.10 indicate that connection strength is by far the most important determiner of the number of perseverations produced, reflecting the key role of connections strength in phoneme reactivation. As for anticipations and for all measures in the previous chapter, the number of steps which pass before selection also has a very strong effect on the number of perseveration errors generated. Finally, the decay rate setting is also important, as this parameter plays a clear part in determining the strength of the influence of the previous production on the current production.

Anticipatory proportion, error rate and non-contextuality of errors

Our overall results showed that whilst different parameter settings allow the spreading activation model to generate both a high proportion of anticipations with a low proportion of perseverations, and a high proportion of perseverations with a low proportions of anticipations, most of the specific models which generate a higher proportion of perseverations and a low proportion of anticipations get excluded when we apply the constraints on error rate and non-contextuality of errors, as defined in section 4.5. We related this result to our finding in the previous chapter that second onset errors have a greater tendency to be non-contextual than first onset errors. We claim that this is because of differences in the way that first onset and second onset contextual errors are generated, such that activation must spread through the network for perseverations to occur. An increase in second onset contextual errors would therefore normally be expected to be accompanied by an

increase in overall error rate as well as an increase in the proportion of errors which are non-contextual.

This finding is reminiscent of a result presented by Dell, Burger, and Svec (1997), who showed that speakers who make more errors also produce a higher proportion of perseverations compared to anticipations. For example, aphasic patients (Schwartz et al., 1994), children (Stemberger, 1989), and people who are under pressure to speak quickly (Dell, 1990; Dell, Burger, & Svec, 1997), make more errors, a smaller proportion of which are anticipations. Conversely, in speakers who make less errors due to practise of a given phrase, the proportion of errors which are anticipations is high (Dell, Burger, & Svec, 1997; Schwartz et al., 1994). Schwartz et al.'s (1994) results further suggested that in the "good" behaviour patterns exhibited by speakers with low error rates and high anticipatory proportions, a higher proportion of the errors resulted in word outcomes.

To measure a speaker's tendency to generate anticipations or perseverations, Dell, Burger, and Svec (1997) calculated the *anticipatory proportion* of the errors generated. This measure is calculated by counting all anticipations and perseverations generated, and determining what proportion of these errors are anticipations. The resulting measure has a minimum of 0 if no anticipations are present and all errors are perseverations, and a maximum of 1 if all errors are anticipations with no perseverations present.

Figure 5.8 shows that anticipatory proportion was also negatively correlated with overall error rate across the specific models in the current data (Kendall's tau = 0.33, $z = 30.20$, $p < 0.001$). The spreading activation model can therefore replicate this pattern of variance in human speakers through manipulations of the spreading activation parameters. Furthermore, in our simulations, a higher anticipatory proportion was also associated with a lower proportion of non-contextual errors, as visible in figure 5.9 (Kendall's tau = 0.57, $z = 50.95$, $p < 0.001$). This finding fits in with Schwartz et al.'s (1994) suggestion that "good" errors occur when error rates are lower, and provides a prediction for which empirical confirmation could be sought, perhaps by comparing normal speakers to groups in which error rate is elevated, or by experimental manipulations of error rate.

Dell, Burger, and Svec (1997) presented a very abstract model where either a past, a present or a future node was selected for production, to help explain this relationship. Their calculations demonstrated that most parameter manipulations which led to decreased error rates also increased the anticipatory proportion. For example,

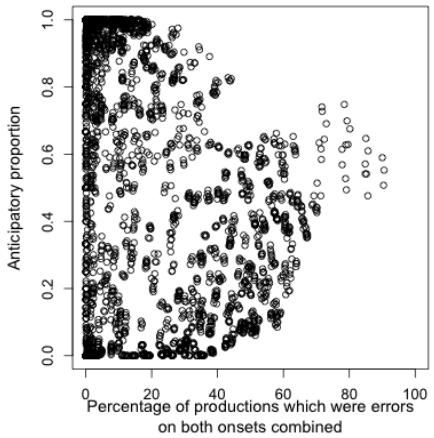


Figure 5.8: Anticipatory proportion plotted against error rate for both onsets combined. Anticipatory proportion can only be calculated for specific models which generated at least one anticipation or perseveration.

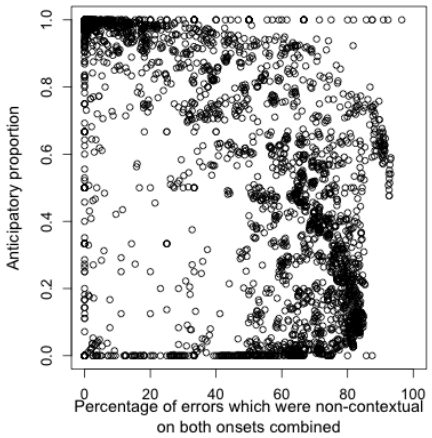


Figure 5.9: Anticipatory proportion plotted against the proportion of errors which were non-contextual on both onsets combined. Anticipatory proportion can only be calculated for specific models which generated at least one anticipation or perseveration.

higher connection strengths, taken to represent better learnt phrases, decreased error rates and increased the anticipatory proportion, as did a higher number of timesteps before output selection, taken to represent a slower speech rate. Higher rates of decay also had this effect by reducing the influence of the previous production and associated perseverative errors. The two parameters which were exceptions to this rule were the amount of activation with which the future node was primed, and a parameter which governed how noisy the decision process was. We therefore examined whether it was possible to relate the effects of parameter manipulations in the current model to the effects of parameter manipulations reported by Dell, Burger, and Svec (1997).

The logistic regression model summarised in table 5.11 and the graphs in figure 5.10 illustrate the effect of parameter manipulations on the anticipatory proportion of errors. Table 5.12 then summarises the directions of effects and Z values for the logistic regressions of parameter effects on anticipatory proportion, error rate and non-contextuality of errors and figure 5.11 compares the medians of these measures at each parameter setting. These tables and graphs further confirm that as in Dell, Burger, and Svec's (1997) model, manipulating parameters such that the anticipatory proportion increases generally resulted in a decrease in error rate and

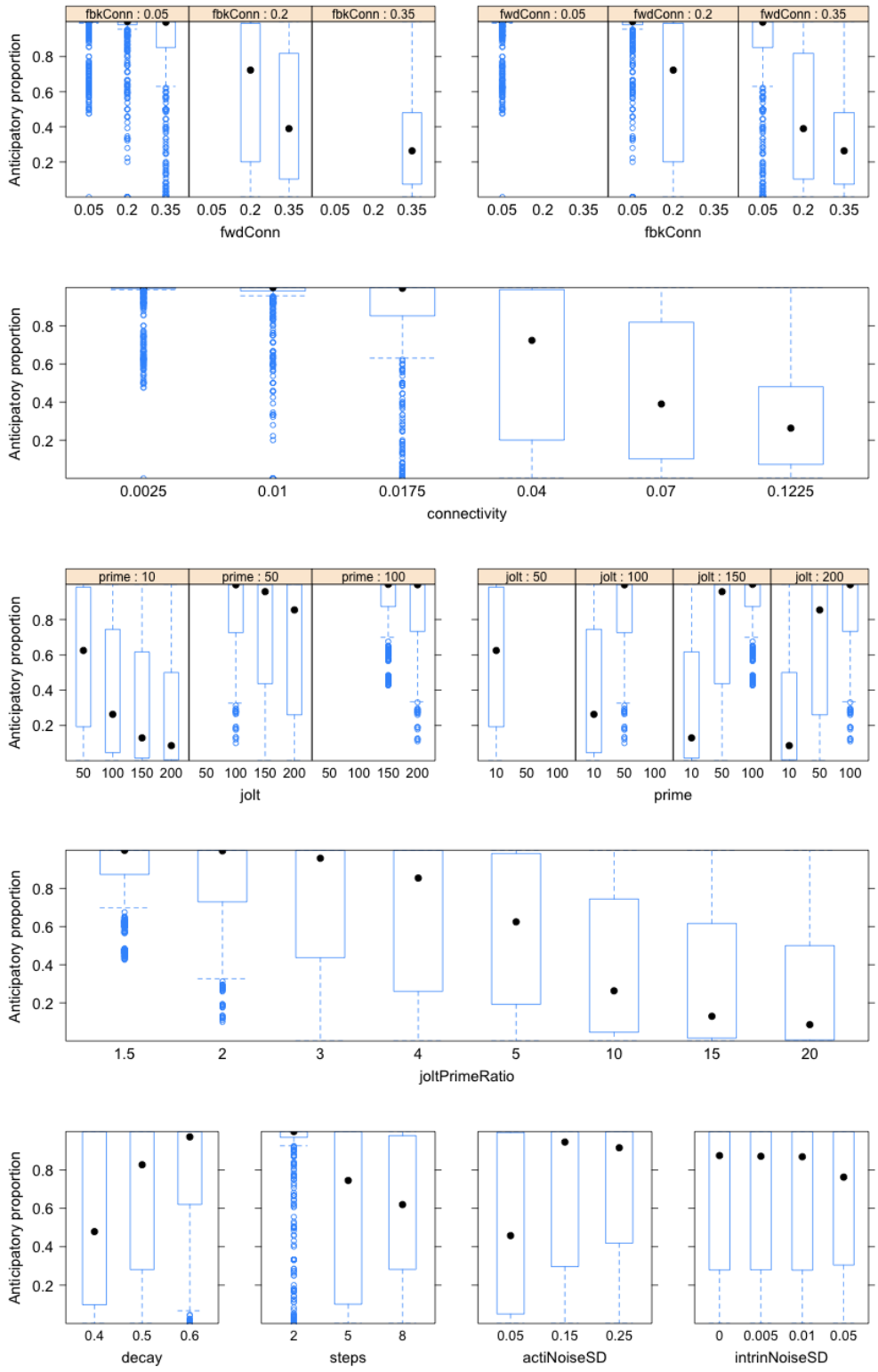


Figure 5.10: The effect of changing parameter settings on the anticipatory proportion. Anticipatory proportion can only be calculated for specific models which generated at least one anticipation or perseveration.

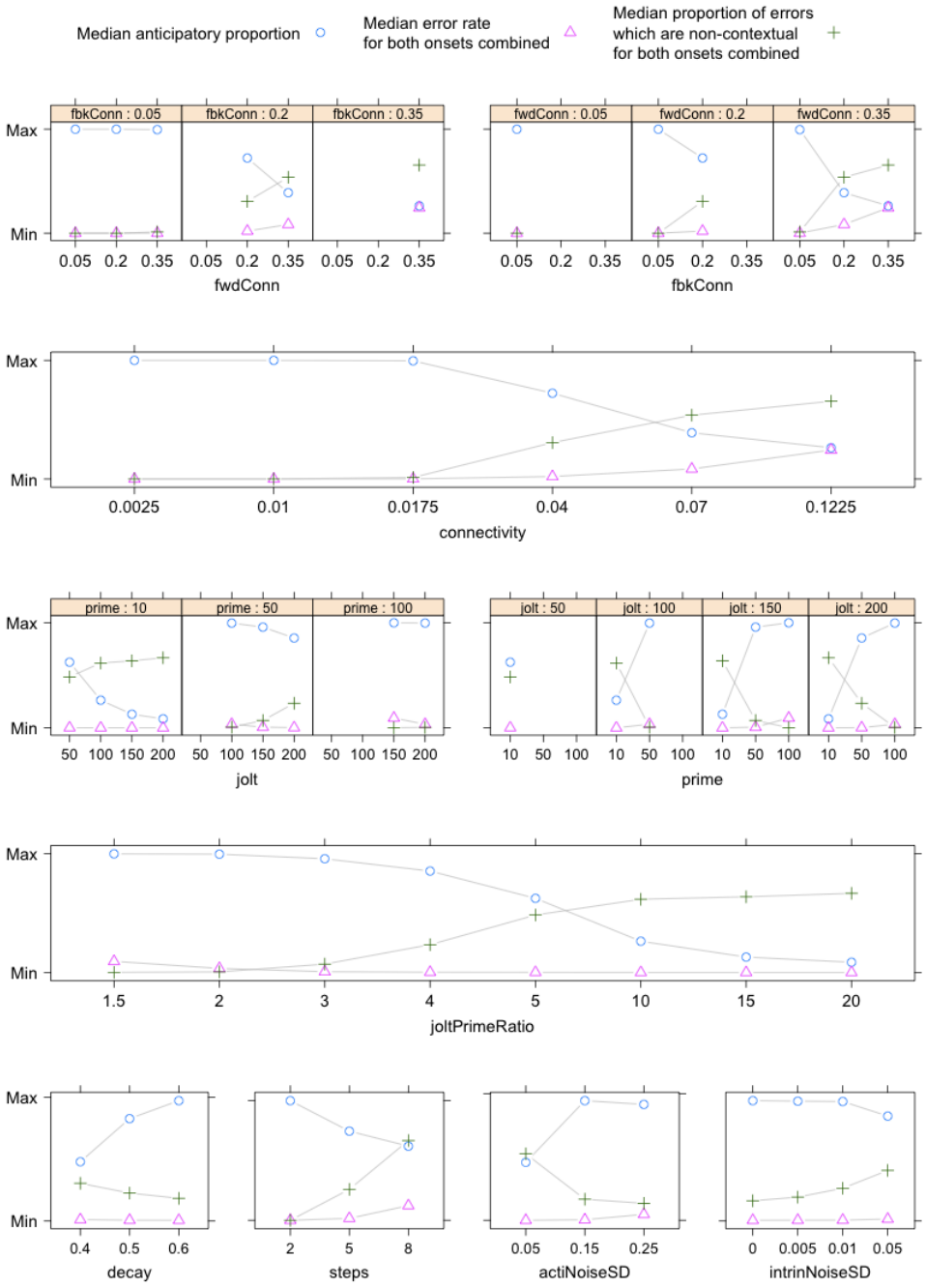


Figure 5.11: The effect of changing parameter settings on the median anticipatory proportion, the median error rate for both onsets combined, and the median proportion of errors which were non-contextual for both onsets combined. To simultaneously show the effect of parameter manipulations on each of these measures, every measure is plotted so that Min on the y-axis marks the lowest possible value for that measure (0 for the anticipatory proportion, and 0% for the error rate and proportion of errors which are non-contextual) while Max on the y-axis denotes the highest possible value for that measure (1 for the anticipatory proportion, and 100% for the error rate and proportion of errors which are non-contextual). The median was chosen as an average measure rather than the mean, as none of these measures are normally distributed. Anticipatory proportion can only be calculated for specific models which generated at least one anticipation or perseveration. The proportion of errors which were non-contextual can only be calculated for specific models which generated at least one error.

Table 5.11: Results of logistic regression model analyses using parameter values to predict the anticipatory proportion of word pair errors. Anticipatory proportion can only be calculated for specific models which generated at least one anticipation or perseveration. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	−	699.5	625150	< .001	*
joltPrimeRatio	−	526.6	557844	< .001	*
decay	+	509.2	299662	< .001	*
steps	−	235.5	59736	< .001	*
actiNoiseSD	+	115.2	13272	< .001	*
intrinNoiseSD	−	10.4	108	< .001	*

Table 5.12: Summary of directions of effects and absolute Wald's Z values from previously reported logistic regression model analyses using parameter values to predict the anticipatory proportion, error rates and proportions of errors which are non-contextual for both onsets combined. Anticipatory proportion can only be calculated for specific models which generated at least one anticipation or perseveration. The proportion of errors which were non-contextual can only be calculated for specific models which generated at least one error.

	Anticipatory proportion		Error rate		Non-contextuality	
Parameter	Direction	Z	Direction	Z	Direction	Z
connectivity	−	699.5	+	2912.5	+	675.3
joltPrimeRatio	−	526.6	−	737.9	+	553.2
decay	+	509.2	−	784.8	−	84.8
steps	−	235.5	+	2864.9	+	887.9
actiNoiseSD	+	115.2	+	1084.1	+	13.9
intrinNoiseSD	−	10.4	+	391.4	+	252.7

the proportion of non-contextual errors generated. The exceptions to this rule were manipulations of the jolt to prime ratio, and the level of activation-based noise. Notably, the prime parameter and a parameter governing how noisy the decision process were also the exceptions in Dell, Burger, and Svec's (1997) model. Decreases in the jolt to prime ratio, which correspond to the prime becoming proportionally bigger in comparison to the jolt, resulted in higher anticipatory proportions, but higher error rates too, although the non-contextuality of errors did decrease. Increases in the level of activation-based noise led to higher anticipatory proportions, higher error rates, and higher proportions of non-contextual errors.

As in Dell, Burger, and Svec's (1997) model, the majority of parameters in the current implementation have a stronger effect on second onset error rate than first onset error rate, and equally affect the number of perseverations generated more than the

number of anticipations generated. Error rate increases caused by these parameters are therefore largely driven by error rate increases on the second onset, causing the negative correlation between error rate and anticipatory proportion. Examples of such parameters which apply in both models are the connection strength, the rate of decay, and the number of steps which pass before selection occurs. As noted earlier in the section, in the current model, errors on the second onset also have a greater tendency to be non-contextual as activation is more dispersed around the network at second onset production, so an increase in second onset error rate also leads to an increase in the proportion of non-contextual errors. The jolt to prime ratio and the level of activation-based noise on the other hand have a much stronger effect on first onset error rate and anticipation generation, such that anticipatory proportion increases with error rate. Whilst increasing activation-based noise also very slightly increases the proportion of non-contextual errors, the jolt to prime ratio has an inverse effect on first onset non-contextuality compared to first onset error rate, such that the negative correlation between anticipatory proportion and non-contextuality of errors is maintained for this parameter. Finally, in the current implementation, the effect of intrinsic noise on error generation at the first and second onset is very similar, and therefore its effect on anticipatory proportion is extremely weak.

The effect of decay on anticipatory proportion and error rate in the current implementation is directly in line with the effect of decay that Dell, Burger, and Svec (1997) describe in their very abstract model, such that a decreased decay rate increases the influence of the previous production on the current production, thereby simultaneously increasing the number of perseverations generated and increasing the error rate.

Other spreading activation model studies which have focused on single word production in normal and aphasic speakers (Dell, Schwartz, et al., 1997; Martin et al., 1994) have elicited higher error rates using higher decay rates, in the same way that the results we reported in section 4.4.2 show that higher decay rates lead to higher error rates on the first onset. As Dell, Burger, and Svec (1997) highlight, it is not clear whether high error rates caused by high decay rates would be accompanied by a decrease in anticipatory proportions.

However, there is a possibility that if a second syllable was produced, which would be necessary to calculate anticipatory proportion, these high decay rates would in fact lead to lower error rates on the second syllable as higher decay rates do here, and therefore lower error rates overall. In our results, the effects of decay

rate manipulations are much stronger for the second onset than for the first onset, such that the decrease in second onset error rate which higher decay rates bring far outweighs the accompanying increase in first onset error rates. The decay rates we tested were however not as high as in the aphasic studies. Whereas we considered behaviour at decay rates of 0.4 to 0.6, Martin et al. (1994) raised decay rates from 0.4 in simulations of normal speakers to 0.92 when simulating aphasic productions, and similarly, Dell, Schwartz, et al. (1997) raised decay rates from 0.5 for normal simulations to a maximum of 0.94 in simulations of one aphasic patient.

Nevertheless, we underline that our results suggest that in models of multiple syllable production, at least with middling decay rates of 0.4 to 0.6, manipulations of decay do allow the spreading activation model to simulate the negative correlation between error rate and anticipatory proportion seen in the empirical evidence reviewed and collected by Dell, Burger, and Svec (1997).

Whilst manipulations of connection strength and the number of timesteps before selection in the current implementation both lead to negative correlations between error rate and anticipatory proportion, as they do in Dell, Burger, and Svec's (1997) more abstract model, the direction of these effects are actually reversed in the current model in comparison to the abstract model.

Dell, Burger, and Svec (1997; see also Dell, 1990) report that speakers talking at higher speeds demonstrate both higher error rates and lower anticipatory proportions. In the abstract model, allowing a higher number of timesteps to pass before selection is taken to represent a slower speech rate, which correspondingly leads to lower error rates and a higher anticipatory proportion, where both results are due to a reduced number of perseverations. We showed in the previous chapter that in the current model, contrary to Dell's (1986) original claims, higher numbers of steps before selection in fact lead to higher error rates, including perseverations. The results reported in this section further demonstrate that anticipatory proportions are lower at higher steps settings.

In the abstract model, when connection strength increases, error rate decreases and anticipatory proportion increases. Dell, Burger, and Svec (1997) suggest that their analysis implies that this should happen in implemented spreading activation models too, such as that presented by Martin et al. (1994), where lower connection strengths are indeed used to cause higher error rates (see also Foygel & Dell, 2000; Dell, Schwartz, et al., 1997) but no analysis of anticipatory proportion is provided. On the whole however, we find that increased connection weights lead

to higher error rates and lower anticipatory proportions. In line with arguments made in earlier sections about the effect of connection weight, this is probably due to the feedback connectivity being strengthened. Higher feedback connection strength supports perseveration generation by facilitating reactivation of previously produced onsets, and also increases the overall error rate by increasing the activation level of representations throughout the network (particularly representations with many connections to other representations), thereby decreasing the influence of the jolt activation. We note that in the abstract model, there is no feedback from output representations to more abstract planning representations, although there is in Martin et al.'s (1994) spreading activation model and subsequent models (e.g., Foygel & Dell, 2000; Dell, Schwartz, et al., 1997).

Importantly however, in the previous chapter we demonstrated that our results show that both connection weights which are too high and connection weights which are too low can cause an increase in error rate and the proportion of non-contextual errors generated. At the connection weights we have tested, the problems of using higher connection weights are more evident than the problems caused by using lower connection weights. Casual examination of our data however shows again that the small set of low connection strength specific models which exhibit high error rates and generate high proportions of non-contextual error rates do indeed display medium rather than high anticipatory proportions. In the same way that low connection strengths cause high error rates by disrupting the effective transmission of the jolt activation, reducing the strength with which the prime is transmitted impacts upon the network's tendency to generate anticipations, which are then frequently replaced by non-contextual errors instead. Unlike in the abstract model therefore, the principal cause of the decrease in anticipatory proportion associated with low connection strengths in this model is not an increase in perseverations, but rather a decrease in anticipations.

To summarise, in line with our finding that many specific models which generate high proportions of perseverations and low proportions of anticipations are excluded by the constraints on error rate and non-contextuality of errors, we have shown in this section that by manipulating the spreading activation parameters of the current model, the empirical pattern reported by Dell, Burger, and Svec (1997) can be simulated, such that specific models which exhibit higher error rates tended to also show lower anticipatory proportions. This aligns the behaviour of the current implementation with the abstract model described by Dell, Burger, and Svec (1997), although there are differences between the way that the parameters of this

model and parameters of Dell, Burger, and Svec's (1997) model affect this negative correlation. Our simulations further demonstrate that in the spreading activation model, high anticipatory proportions are associated with lower proportions of non-contextual errors, a result which could be linked by to Schwartz et al.'s (1994) suggestions that "good" errors occur when anticipatory proportions are higher and error rates lower. This provides a prediction for which empirical support could be sought.

Anticipation and perseveration results summary

Our results show that very few specific models simultaneously generated appropriate proportions of anticipations and appropriate proportions of perseverations according to the new bounds we determined in section 5.2.1. The proportional-incompletes bounds proved to be a particular problem for the model, to the extent that when we excluded specific models which generated either too many errors or too high a proportion of non-contextual errors for the limits established in section 4.5, no specific models met these bounds at all. However, it was highlighted that the model's inability to behave within these bounds may be due to a potentially very low occurrence rate of exchange errors, a suggestion which is investigated further in the following section.

Instead, we found that large numbers of specific models generated too many anticipations and not enough perseverations, whilst many other specific models generated too many perseverations and not enough anticipations. However, applying the constraints on error rate and non-contextuality of errors led to many of the models which generated too high a proportion of perseverations being excluded, whereas many of the models which generated too high a proportion of anticipations remained. It was noted that this result was in line with the finding in the previous chapter that high proportions of second onset errors are non-contextual. This means that an increase in the number of second onset contextual errors is likely to be accompanied by a large increase in the number of second onset non-contextual errors, leading to increases in both the overall error rate and the overall proportion of non-contextual errors.

Effects of spreading activation parameter manipulations on anticipation and perseveration generation were found to largely reflect the effects of parameter manipulations on first and second onset error generation respectively. The jolt to prime ratio and level of activation-based noise were particularly important in determining the number of anticipations generated, whereas the connection strength, number of

timesteps before selection, and decay rate were of most importance in determining how many perseverations occurred. The key difference between the effects of parameters on second onset error rate and on perseveration generation was that a high jolt to prime ratio led to an increase in perseverations, contrary to the decrease in overall second onset errors that this manipulation causes. However, this result fits in with our finding in the previous chapter that a higher jolt to prime ratio leads to a higher proportion of contextual errors on the second onset, and the explanation provided as to why this parameter setting leads to more perseverations.

In line with these results, we showed that increases in error rate caused by manipulations in connection strength, the number of steps before selection, or decay rate are associated with lower anticipatory proportions, demonstrating that the empirical negative correlation between these two measures which was reported by Dell, Burger, and Svec (1997) can be simulated in the spreading activation model. Manipulations of jolt to prime ratio and activation-based noise do not support this relationship however, as error rate increases associated with these parameters lead to more anticipations than perseverations. These results fit in with the behaviour of the very abstract model that Dell, Burger, and Svec (1997) presented to explain this effect, where manipulations of connection strength, the number of steps before selection, or decay rate cause variation in which there is a negative correlation between anticipatory proportion and error rate, but manipulations of prime and the noise affecting the selection process do not.

We added to these results by demonstrating that decay manipulations in a spreading activation model can in fact cause this correlation between anticipatory proportion and error rate, as long as low rather than high decay rate is the cause of the increase in error rate, unlike in Martin et al.'s (1994) and later Dell, Schwartz, et al.'s (1997) model. We also demonstrated that in the current model, high connection strength and high numbers of steps before selection cause low anticipatory proportions and the associated high error rates, rather than the low connection strengths and low numbers of steps which were responsible for this effect in Dell, Burger, and Svec's (1997) model.

Lastly, we showed that in the current model, higher anticipatory proportions are associated with lower proportions of non-contextual errors, a result which fits in with Schwartz et al.'s (1994) suggestion that "good" errors (such as errors which result in word outcomes) occur at low error rates. This finding provides a prediction which could be tested empirically, by considering groups in whom error rate is naturally elevated, or by experimentally manipulating error rate.

5.4.2 Exchange errors

We finally turn to examine the implementation's ability to generate appropriate proportions of exchange errors, and the influence of parameter settings on this behaviour.

Overview of implementation behaviour

Again, we began by considering all 4092 specific models which generated contextual errors for analysis, 70.2% of the 5832 models tested. As suspected from our examination of Dell's (1986) original simulation results reported in section 5.2.2, and the anticipation and perseveration results reported in the previous section, the model showed particular difficulty with exchange error generation. Figures 5.12 and 5.13 show that very few specific models generated a sufficiently high proportion of exchange errors. For the primary proportional-incompletes analysis, 4075 specific models generated too few exchanges (69.9% of all models tested), and only 17 specific models generated over the 14.3% exchange errors required (0.3% of all models tested). Of these specific models, 12 also generated proportions of anticipations and perseverations which fell within the specified bounds (0.2% of all models tested). The remaining 5 specific models (0.1% of all models tested) generated an appropriate proportion of perseverations, but too many anticipations.

For the secondary incompletes-as-anticipations analysis, which requires much fewer exchanges, 3877 specific models generated relatively too few exchanges (66.5% of all models tested), with only 212 generating over the 5% exchange errors required (3.6% of all models tested). These 212 specific models included 85 which also passed both the anticipation and perseveration constraints (1.5% of all models tested), 101 which generated too many perseverations (1.7% of all models) and 26 which generated too many anticipations (0.4% of all models). The final 3 specific models (0.1% of all models tested) actually generated too many exchanges, but an appropriate proportion of anticipations and perseverations.

When considering all models which generated errors for analysis, there were therefore specific models which passed the anticipation, perseveration and exchange constraints simultaneously for both analyses, but very few of them, with only 12 specific models for the proportional-incompletes analysis (0.2% of all models tested), and 85 (1.5% of all models tested) for the secondary incompletes-as-anticipations analysis.

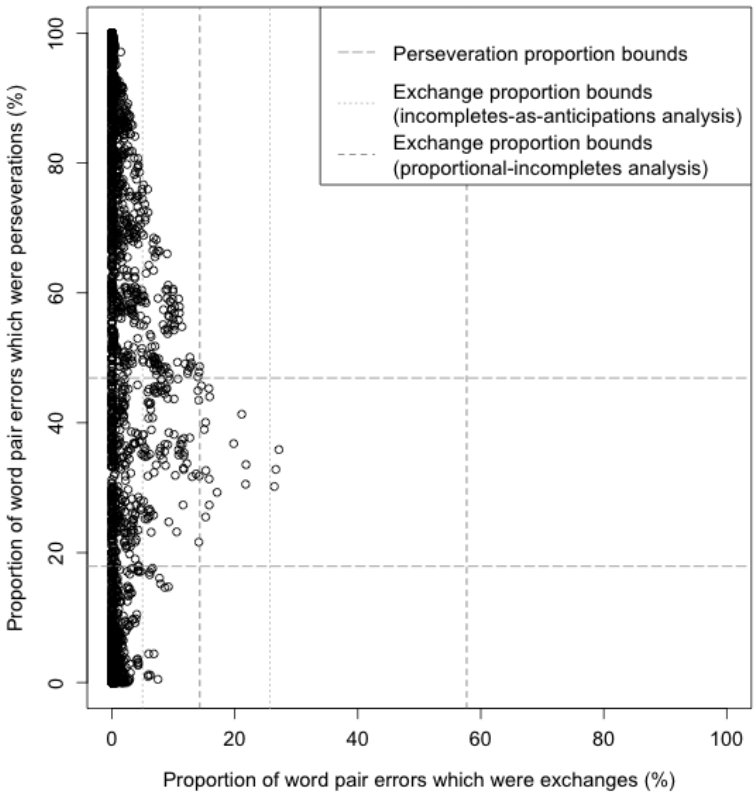


Figure 5.12: The proportion of contextual word pair errors which were exchanges plotted against the proportion of contextual word pair errors which were perseverations, for all specific models which generated at least one anticipation, perseveration or exchange

The situation looks even worse when only the 1665 specific models which generated sufficiently few errors and sufficiently low error rates are considered (28.5% of all models tested). Figures 5.14 and 5.15 show that all 1665 simulations generate relatively too few exchanges for the primary proportional-incompletes analysis bounds, and only 2 simulations generate enough exchanges to fall within the bounds specified by the secondary incompletes-as-anticipations analysis (0.03% of all models tested). It can clearly be seen from the figures that both of these two simulations generate too few perseverations, and too many anticipations.

Firstly, we note that these results confirm that exchange production caused a big problem for the specific models when trying to meet the bounds of appropriate anticipation and perseveration generation. With so few exchanges generated, anticipation and perseveration proportions were inflated such that it was impossible for

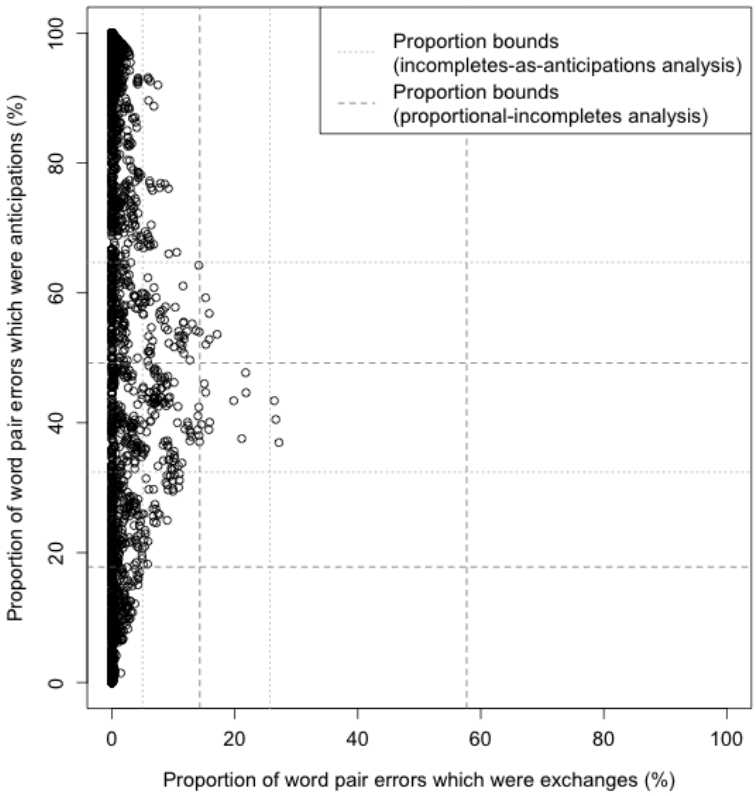


Figure 5.13: The proportion of contextual word pair errors which were exchanges plotted against the proportion of contextual word pair errors which were anticipations, for all specific models which generated at least one anticipation, perseveration or exchange

many models to meet the proportional-incompletes anticipation and perseveration bounds in particular.

Secondly, we argue that these results suggest that the trigger mechanism proposed by Dell (1986) to generate errors on the second onset when an error has occurred on the first onset is not strong enough. To recap, given the target production *"big fun"*, the exchange error *"fig bun"* is proposed to occur as follows. Firstly, the anticipatory error *"fig"* is produced as in an anticipation, due to the influence of priming activation applied to the /f/ onset and noise in the network, which together cause the /f/ phoneme to have more activation than the target /b/ phoneme. Selection of the /f/ onset for production causes its activation to be completely suppressed. As the intended /b/ onset has not been selected, its activation is not suppressed such that some activation should remain on this node. The suggestion made by Dell (1986) is that this activation, combined with noise in the network, should then

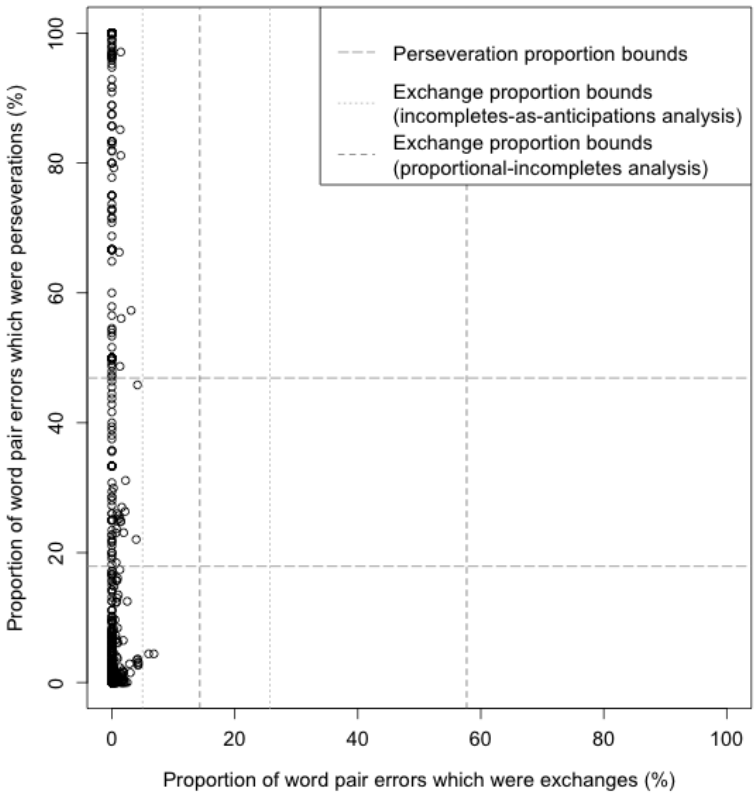


Figure 5.14: The proportion of contextual word pair errors which were exchanges plotted against the proportion of contextual word pair errors which were perseverations, for all specific models which passed both constraints on overall error rate and non-contextuality of errors

at second onset production lead the /b/ phoneme to be more activated than the target /f/ phoneme, triggering the completion of the exchange, *"fig bun"*.

Our results suggest however that the activation remaining on the intended first onset is not enough to trigger the second part of the exchange on a frequent enough basis. Indeed, the activation on the intended but unselected first onset /b/ has to compete with the large jolt activation passed to the intended second onset /f/. In addition, as the /f/ node was selected as the first onset due to being the most activated onset phoneme, it is also quite possible that other words beginning with /f/, such as *fill* and *fat*, have received a substantial amount of activation during first onset production, which will be transmitted back to the /f/ onset during second onset production. In specific models where the exchange proportion is higher, we argue that this is largely caused by the sheer force of a higher error rate overall, which will increase the proportion of coincidentally occurring double errors (i.e., exchanges) in

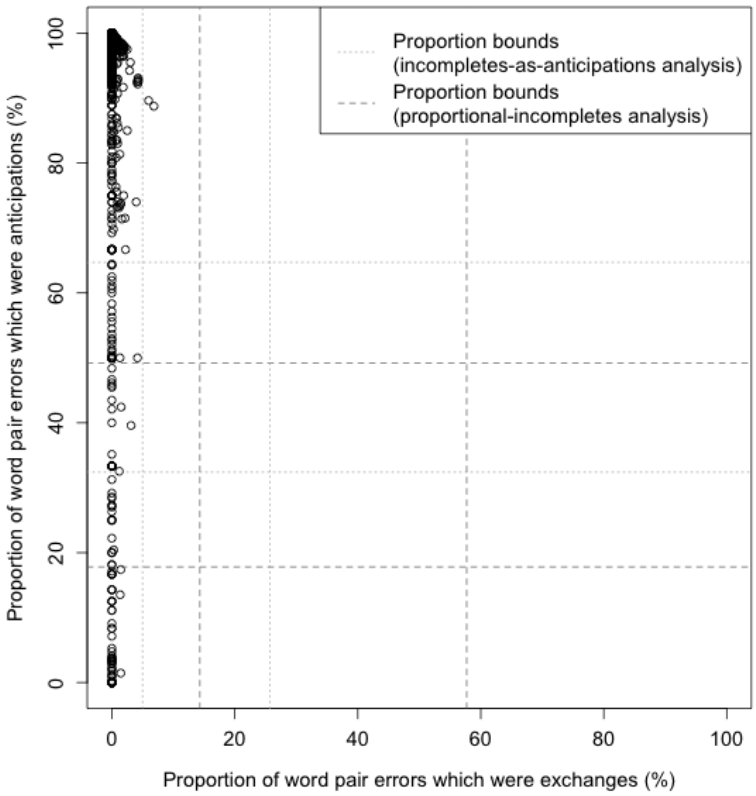


Figure 5.15: The proportion of contextual word pair errors which were exchanges plotted against the proportion of contextual word pair errors which were anticipations, for all specific models which passed both constraints on overall error rate and non-contextuality of errors

comparison to the proportion of single errors accompanied by a correct production (i.e., anticipations and perseverations). As we have repeatedly highlighted however, human speakers do maintain some degree of accuracy. A satisfactory model of exchange generation therefore cannot rely on exceedingly high error rates. These models are therefore excluded when we apply the constraint on error rate which we determined in chapter 4.

The next sections will look at the effects of manipulating parameters on exchange generation, to try to determine what chance there would be of finding parameter settings outside the current parameter space at which the model would generate enough exchanges. While we already know that most of the models which pass the constraints on error rate and non-contextuality generate too many anticipations and too few perseverations, we also investigate whether the results of our parameter explorations can throw any further light on why the two models which do not

generate too many errors but do generate sufficient exchanges exhibit this particular pattern.

Effects of parameter manipulations on exchange error generation

The logistic regression summarised in table 5.13 and the graphs in figure 5.16 depict the effects of manipulating the spreading activation parameters on exchange error rates. Again, these are percentages of all word pair productions that are exchanges, rather than proportions of all word pair errors, as this latter measure would be heavily influenced by anticipation and perseveration generation. The previous chapter showed that with the exception of manipulations of decay rate, most parameter manipulations which led to an increase in first onset errors also led to an increase in second onset errors, even if the size of this increase differed. The directions of the effects of these parameter manipulations on exchange error generation therefore mimic the directions of the effects of these parameter manipulations on first and second onset error generation, with high connection strength, low jolt to prime ratios, high numbers of steps before selection, and high levels of activation-based and intrinsic noise all leading to higher exchange error rates. As for second onset error generation, lower decay rates lead to higher exchange error rates, as the effect of decay rate is much stronger on second onset error rates than first onset error rates.

The explanation of these effects is quite straightforward for most parameters, though a little more complicated for connectivity strength and jolt to prime ratio. A low decay rate boosts exchange error generation as it helps maintain the activation of the unselected first onset. A high number of steps before selection causes more errors on both onsets by reducing the influence of the jolt activation and permitting noise to have a greater effect on activation levels. Increasing either type of noise also reduces the influence of the jolt and increases error rates, although activation-based noise is particularly important as its effects on the primed onset are so strong.

High connectivity strength causes more errors on both onsets, though particularly the second onset. However, we note that while the role of connectivity in reactivating phonemes to elicit perseverations is very clear, a different mechanism is responsible for generating exchange errors. Whilst the overall increase in error rate caused by increases in connectivity strength will increase the coincidental occurrence of exchange errors as explained previously, we suggest that the increased activation levels in the network caused by higher connectivity strengths perhaps allow the activation on the intended first onset to remain at a higher level such that it can compete more effectively with the jolted target second onset. Furthermore,

Table 5.13: Results of logistic regression model analyses using parameter values to predict the percentage of word pair productions which were exchanges. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	+	205.2	44424	< .001	*
joltPrimeRatio	−	125.4	43620	< .001	*
decay	−	109.3	13057	< .001	*
steps	+	154.2	29447	< .001	*
actiNoiseSD	+	188.8	58720	< .001	*
intrinNoiseSD	+	14.3	201	< .001	*

we again note that figure 5.16 indicates that reducing the connectivity strength to the very lowest setting causes a small rise in exchange errors for some specific models. We suggest that this is entirely due to the very high error rate exhibited by these specific models, as demonstrated in the previous chapter.

Finally, low jolt to prime ratios lead to more errors on the first onset because the upcoming onset is more strongly primed. However, they also make exchanges more likely on the second onset, as we argued in section 4.4.3. When the prime is high relative to the jolt, anticipations on the first onset are more likely, which leads to the intended first onset not being reset. Once an anticipation has occurred, and the intended second onset has been reset, a low jolt to prime ratio also means that less jolt activation is provided to the intended second onset in proportion to the activation already in the network. Altogether, this means that the probability that the activation remaining on the intended first onset is more than the activation on the intended second onset is therefore increased, leading to more completed exchange errors.

The difference in importance of the effects of different parameter manipulations is less for exchange error generation than it is for anticipation and perseveration generation, and we do not comment further on this here.

The parameters of the two models which pass both the incompletes-as-anticipations exchange error lower bounds and the constraints on error rate and non-contextuality strongly reflect these findings. For these two specific models, connectivity strengths were high (0.1225; i.e., forward connectivity strength was 0.35 and feedback connectivity strength was 0.35), decay rates were low (0.4), activation-based noise levels

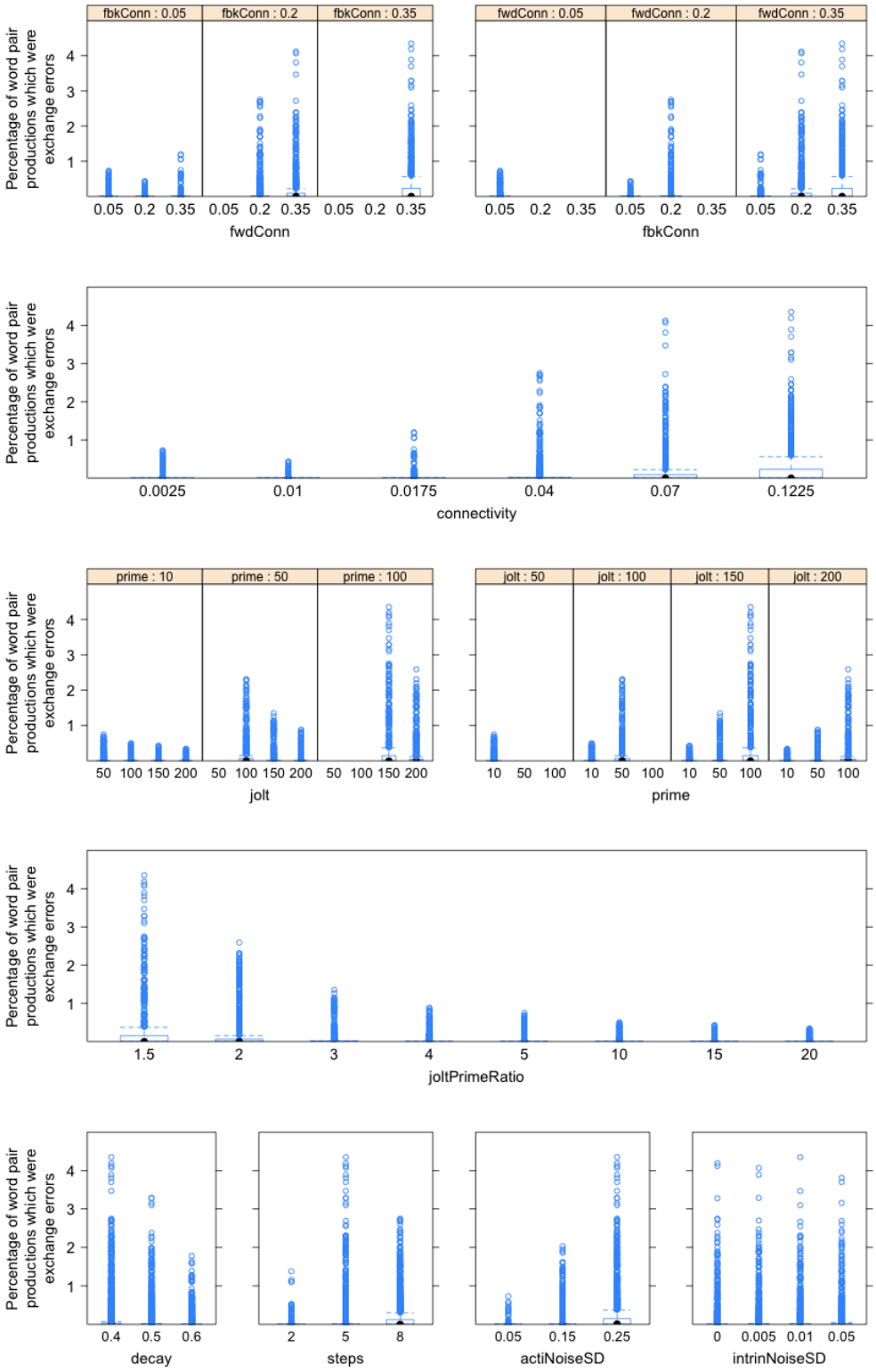


Figure 5.16: The effect of changing parameter settings on the percentage of word pair productions in each simulation which were exchange errors, for all specific models. Note that the y-axis in this graph uses a different scale to figures 5.6 and 5.7, as the model generates fewer exchanges than anticipations and perseverations.

were high (0.25), and intrinsic noise levels apparently unimportant, with one specific model using the lowest setting (0) and the other the highest setting (0.05). However, whilst the jolt to prime ratio was low at 2 (such that jolt was 200 and prime was 100), it was not the lowest. Further investigation of our results showed that with the activation-based noise level set to 0.25, the lowest jolt to prime ratio (1.5, with jolt at 150 and prime at 100) generated too many errors, as both high activation-based noise levels and a low jolt to prime ratio lead to an increase of errors on the first onset. Additionally, the number of steps before selection stage were set to the lowest possible setting (2), rather than the highest. This was perhaps predictable, as we had previously noted that increasing the number of steps was a major cause of non-contextual errors, especially on the second onset. With the connectivity strength set to its highest value, no specific model was able to pass the constraints on erroneousness if there were more than 2 steps per selection stage, with nearly all such specific models generating too many non-contextual errors.

Having noted that increasing the number of steps to 5 or 8 created so many problems in terms of non-contextual errors, and that intrinsic noise had little effect on exchange error generation, we explored our simulation data to uncover the effects of individually varying the four remaining parameters (activation-based noise levels, jolt to prime ratio, connectivity strength and decay rate) whilst holding all parameters apart from the parameter under examination at their optimal settings, with the exception of intrinsic noise which was allowed to vary freely at all times. We examined first and second onset error rates, and anticipation, perseveration and exchange percentages, to search for indications of parameter manipulations beyond our parameter space which could increase the triggering tendency of the model and solve the exchange error generation problem.

Of particular interest is figure 5.18. This figure further confirms our assertion that low jolt to prime ratios support exchange error generation, whilst high jolt to prime ratios support perseveration generation. This also probably explains why the proportion of perseverations is so low for these specific models which generate an appropriate amount of exchange errors at a reasonable error rate.

However, it does not look likely that a manipulation of the jolt to prime ratio will solve the exchange error generation problem. At a lower jolt to prime ratio, the number of exchanges would be higher, but the number of anticipations would also be higher. This would cause the model to generate too many errors, as demonstrated by the exclusion of models with a jolt to prime ratio of 1.5 when our constraints on error rate are applied, as well as causing more anticipations for the exchanges

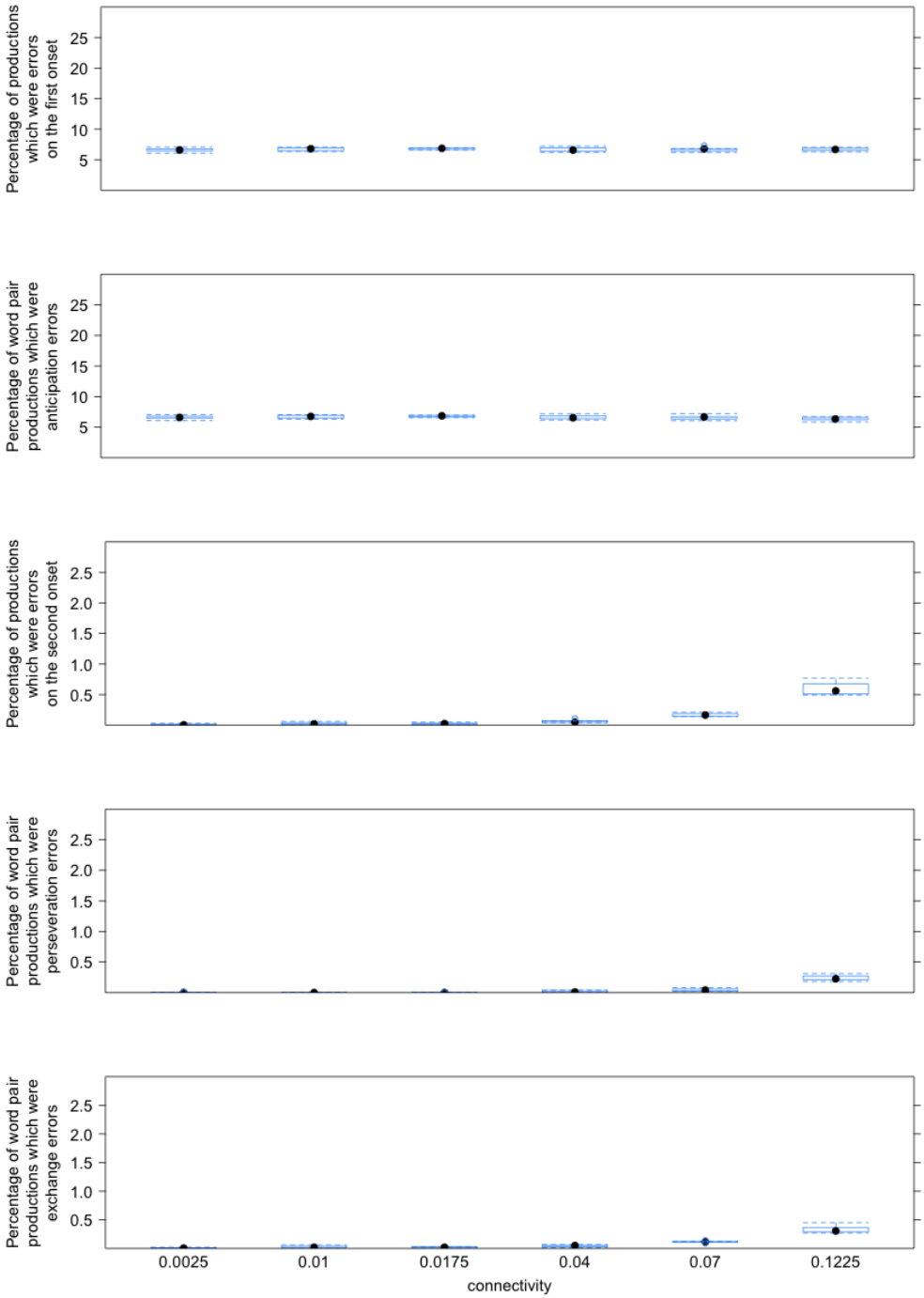


Figure 5.17: The effect of changing the connectivity strength on the first onset error rate, anticipation rate, second onset error rate, perseveration rate and exchange rate, for specific models with a low jolt to prime ratio (joltPrimeRatio = 2; i.e., jolt = 200 and prime = 100, or jolt = 100 and prime = 50), low decay (decay = 0.4), high activation-based noise level (actiNoiseSD = 0.25), a low number of steps per selection stage (steps = 2) and a high activation-based noise level (actiNoiseSD = 0.25). For clarity purposes, the y-axis is scaled differently for the first onset error rate and anticipation error rate than for the other graphs.

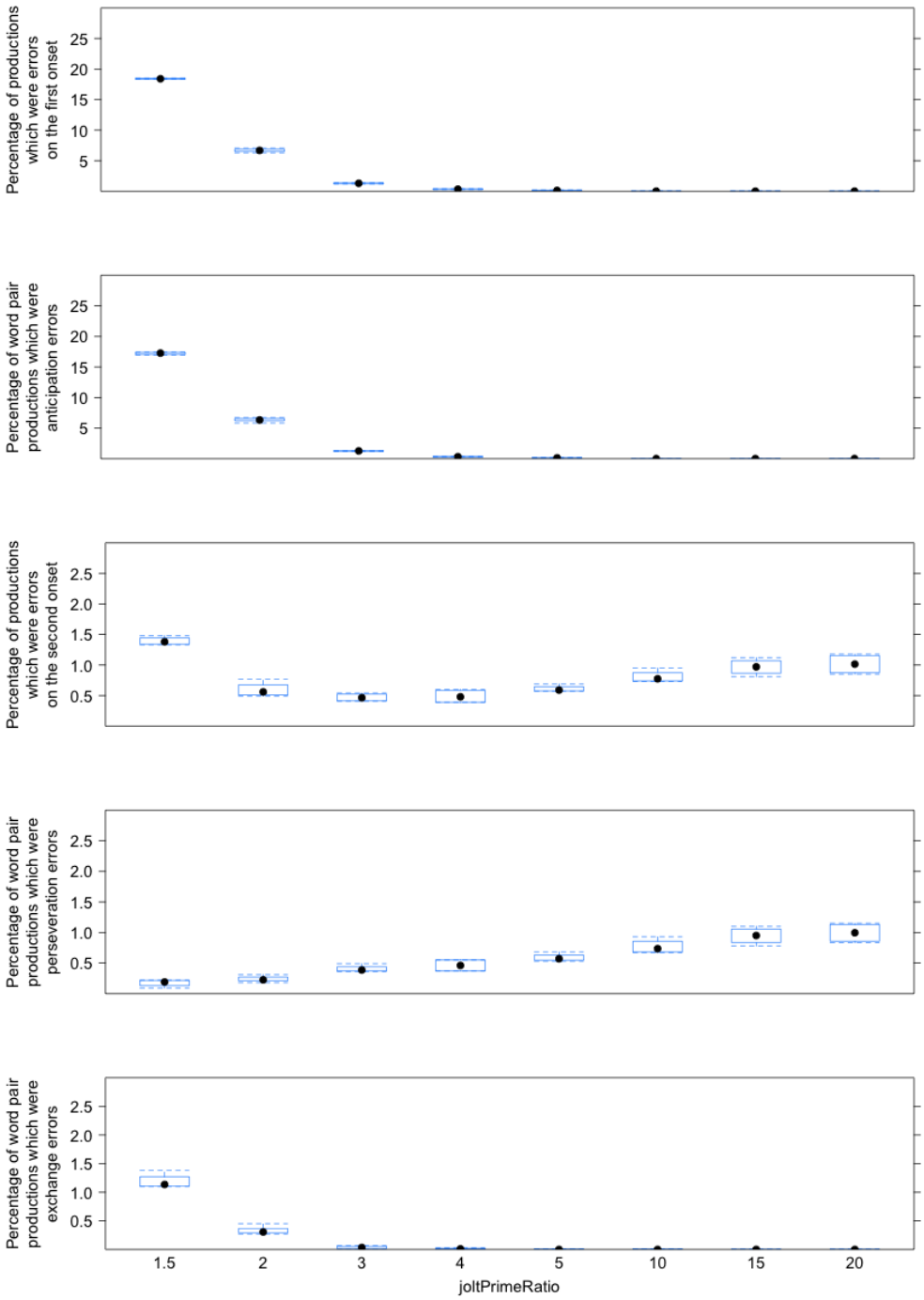


Figure 5.18: The effect of changing the jolt to prime ratio on the first onset error rate, anticipation rate, second onset error rate, perseveration rate and exchange rate, for specific models with high connectivity (connectivity = 0.1225; i.e., fwdConn = 0.35 and fbkConn = 0.35), low decay (decay = 0.4), a low number of steps per selection stage (steps = 2) and a high activation-based noise level (actiNoiseSD = 0.25). For clarity purposes, the y-axis is scaled differently for the first onset error rate and anticipation error rate than for the other graphs.

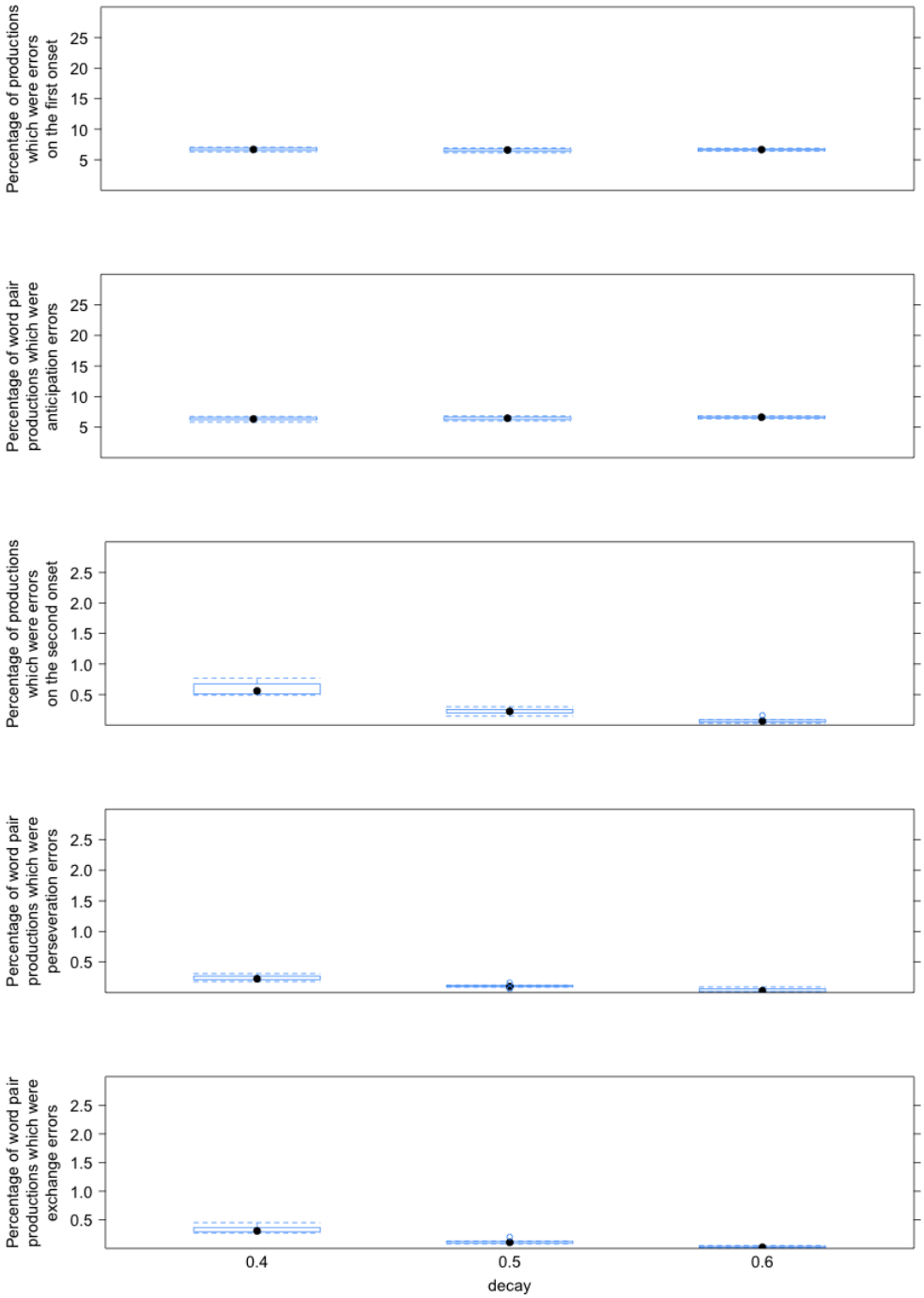


Figure 5.19: The effect of changing the decay rate parameter on the first onset error rate, anticipation rate, second onset error rate, perseveration rate and exchange rate, for simulations with high connectivity (connectivity = 0.1225; i.e., fwdConn = 0.35 and fbkConn = 0.35), a low jolt to prime ratio (joltPrimeRatio = 2; i.e., jolt = 200 and prime = 100, or jolt = 100 and prime = 50), a low number of steps per selection stage (steps = 2) and a high activation-based noise level (actiNoiseSD = 0.25). For clarity purposes, the y-axis is scaled differently for the first onset error rate and anticipation error rate than for the other graphs.

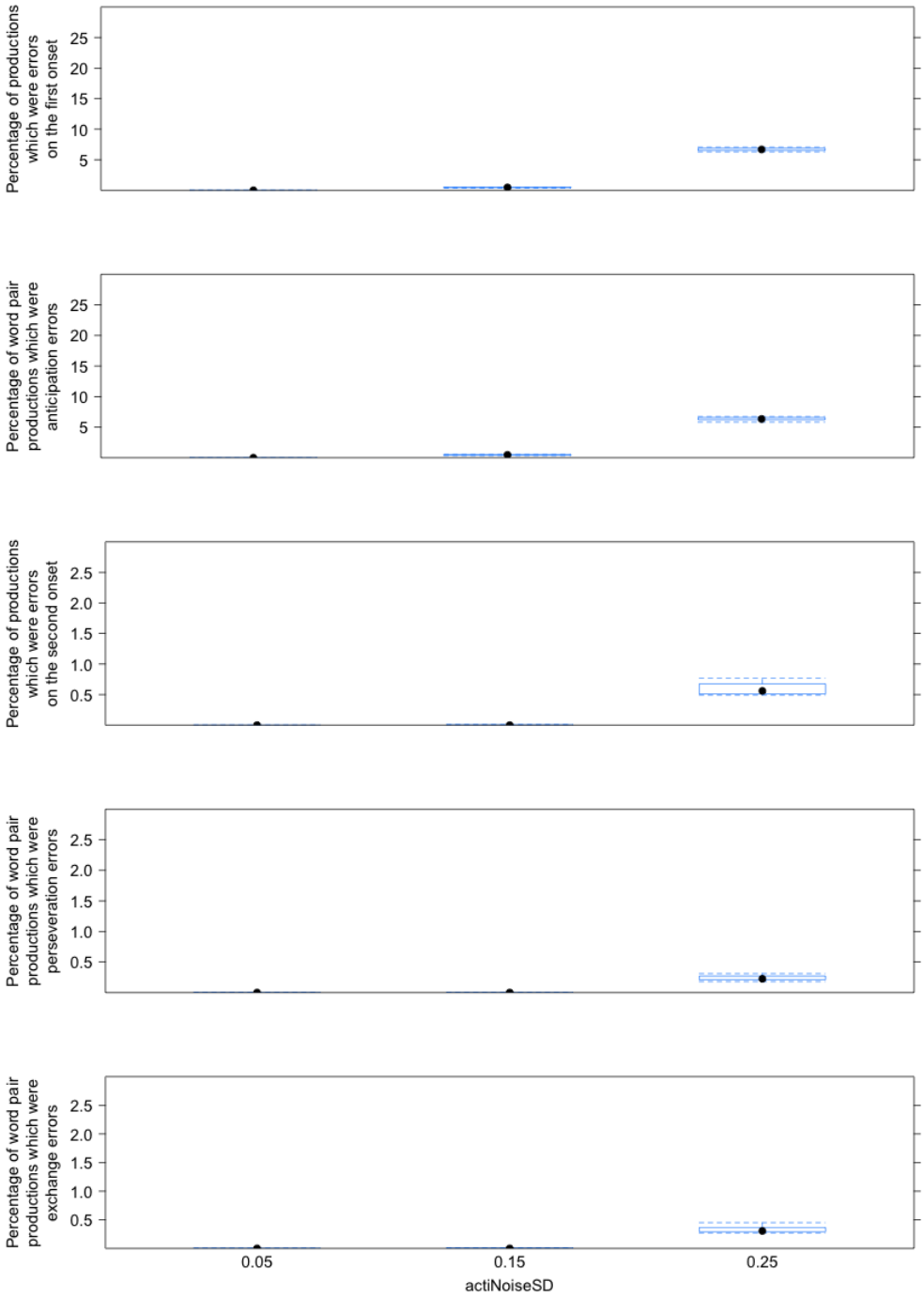


Figure 5.20: The effect of changing the activation-based noise level parameter on the first onset error rate, anticipation rate, second onset error rate, perseveration rate and exchange rate, for simulations with high connectivity (connectivity = 0.1225; i.e., fwdConn = 0.35 and fbkConn = 0.35), a low jolt to prime ratio (joltPrimeRatio = 2; i.e., jolt = 200 and prime = 100, or jolt = 100 and prime = 50), low decay (decay = 0.4) and a low number of steps per selection stage (steps = 2). For clarity purposes, the y-axis is scaled differently for the first onset error rate and anticipation error rate than for the other graphs.

to compete with. At higher jolt to prime ratios, the model's tendency to generate exchange errors would be reduced.

The figures suggest that instead, the model's best hopes would lie in an increase of connection strength, and a decrease in decay rate, taking both of these parameter settings outside the values used in the literature so far, with the exception of a single study reported by Dell (1990) where a decay rate of 0.2 is used. The graphs show that these manipulations would increase second onset error rate, boosting perseveration and exchange error generation, whilst having little effect on first onset error rates. However, the rise in second onset error rate would cause the overall error rate to rise as well. More importantly, Shrager et al. (1987) have demonstrated that connection strength must be lower than decay rate to prevent activation levels rising without bound. Given that our best current parameter settings use a decay rate of 0.4 and a connection weight of 0.35, there is clearly limited room for improvement in this direction. Most critically, there is no clear parameter manipulation option for increasing the triggering tendency of the model rather than simply increasing the error rate.

In summary, whilst further parameter explorations could attempt to resolve the exchange error generation problem, it appears unlikely that they would be successful. Furthermore, the current results plainly demonstrate that a large part of the parameter space for the spreading activation model, covering nearly all of the space explored by previous studies of Dell's (1986) twenty year old model, leaves the model distinctly unable to account for corpus movement error evidence.

Exchange error results summary

Across all 5832 parameter settings tested in the current investigation, only 17 models (0.3% of all models tested) generated more than the 14.3% exchanges required to meet the proportional-incompletes proportion bounds, and only 215 models (3.7% of all models tested) generated more than the 5% exchange errors required by the incompletes-as-anticipations proportion bounds. Within this small group of models, some also met the anticipation and perseveration bounds, with 12 successful specific models (0.2% of all models tested) for the proportional-incompletes analysis, and 85 (1.5% of all models tested) for the secondary incompletes-as-anticipations analysis.

However, once specific models which failed the constraints on error rate and non-contextuality of errors were excluded, no specific models generated over 14.3% exchange errors, such that all models failed the proportional-incompletes analysis, and only 2 (0.03% of all models tested) generated more than 5% exchange errors, as required to pass the lower bound on exchange error proportions set by the incompletes-as-anticipations analysis. However, these models generated too many anticipations and too few perseverations.

We note that the fact that the model generates such low proportions of exchanges overall also effectively prevents specific models from being able to simultaneously generate low enough proportions of anticipations and perseverations to meet both the anticipation and perseveration bounds.

The model's difficulties clearly cannot solely be explained by the classification of some incomplete errors as exchanges. Whilst the behaviour of the model is far off the bounds defined by the proportional-incompletes analysis, even with all incomplete errors classified as anticipations, only two models exhibit exchange error proportions which are barely above the incompletes-as-anticipations lower bound. Instead, we suggest that the trigger mechanism in the model is not strong enough, such that the amount of activation left on the first onset is not enough to cause anticipations to turn into exchanges on a frequent enough basis. Models which do generate enough exchanges do this solely because they exhibit very high error rates overall, such that the probability of coincidentally occurring double errors (exchanges) increases, whilst the probability of single errors accompanied by a correct production (i.e., anticipations and perseverations) decreases. As humans do maintain some level of accuracy in their speech, models which exhibit very high error rates are inappropriate, and are correspondingly excluded by our constraints on error rate and non-contextuality of errors.

Investigations of the effects of parameter manipulations on exchange error generation confirmed that low jolt to prime ratios are better for exchange generation, whereas perseveration generation is more successful at high jolt to prime ratios. However, further manipulations of jolt to prime ratio are unlikely to permit the model to capture the empirical evidence, as low jolt to prime ratios also cause many anticipations and therefore a very high error rate. Our parameter explorations did not uncover any parameter manipulation which would increase the triggering strength of the network. Higher connection strengths and lower decay rates may

help by increasing second onset error rate. However, only limited further manipulation of these parameters would be possible, without activation level representations in the network rising without bound.

It is therefore not clear that these manipulations would be sufficient to close the gap between the model's current behaviour and the empirical benchmarks. Furthermore, regardless of how likely it is that further parameter exploration would lead us to a particular set of parameters at which the model could generate a higher proportion of exchanges, the current investigation shows that across a very large portion of the parameter space, including nearly all of the parameter space considered by previous studies, the spreading activation model cannot account for exchange error evidence.

Finally, the possible difference in implementation in this model and Dell's (1986) model as outlined in section 4.2.3 should be addressed. Whereas in this model, the activation of the first jolted word is reset following selection of the phonemes in the first word, it is not clear that Dell (1986) reset word activation in this fashion. Not resetting activation of the word would be likely to increase second onset error rate, but would not increase the strength of the trigger mechanism, as this extra activation would be present whether the first onset was erroneously produced or not. As previously noted, an overall increase in error rate may simply lead to models being excluded on the basis of the error rate constraints. Moreover, we note that even if Dell (1986) did not reset activation of previously produced words, we showed in section 5.2.2 that his simulations did not in fact exhibit appropriate behaviour either.

5.5 Conclusions

In this chapter, we re-evaluated corpus estimates of the relative rates of anticipation, perseveration and exchange error generation, using multiple speech error corpora, and considering carefully the classification of incomplete errors, such as *"big fun"* → *"fig... big fun"*. We then re-evaluated the behaviour of Dell's (1986) spreading activation model in the light of these revisited corpus analyses, investigating model behaviour across a range of spreading activation parameter settings, whilst also taking into account the limits on error rate and non-contextuality of errors as determined in section 4.5. This section summarises our findings.

5.5.1 Re-evaluation of behavioural evidence

In our re-evaluation of the empirical evidence, we identified four speech error corpus reports where the number of incomplete errors was explicitly reported (del Viso et al., 1991; Nooteboom, 2005b; Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1989). Incomplete errors formed substantial portions of these reports; between a quarter and half of the errors recorded. However, we noted that even without considering the influence of incomplete errors, proportions of anticipations, perseverations and exchanges varied greatly between the corpora.

We presented two analyses of the data in these corpus reports. In the primary analysis, the *proportional-incompletes* analysis, some incomplete errors were classified as anticipations, and others were classified as exchanges. In the secondary analysis, the *incompletes-as-anticipations* analysis, all incomplete errors were categorised as anticipations, as in Nooteboom's (1969) original data. Neither of these analyses suggested that it is true across all corpora that anticipations occur more frequently than perseverations, and perseverations occur more frequently than exchanges. In fact, no ordering of the error categories provided an appropriate generalisation of the data. Instead, we calculated very liberal bounds on the proportions of individual error categories, looking at the most extreme proportions reported across all the corpora, and also considering a one standard deviation bound around the mean, and finally picking the most extreme upper and lower bounds from these calculations.

We then showed that Dell's (1986) original simulation results do not meet these newly determined bounds. In particular, the model appeared to struggle with exchange error generation, such that too few were generated. No results reported by Dell's (1986) met the bounds from the primary proportional-incompletes analysis, as exchange error proportions were too low. The only parameter setting which permitted the model to generate enough exchanges for the much lower bound determined from the secondary incompletes-as-anticipations analysis was accompanied by an extremely high error rate of 16.9%, far above the upper limit on error rate determined in section 4.5 of 5.75%. As the proportion of exchanges generated by chance should rise as the error rate rises, this result suggested that the model was relying on inappropriately high error rates to generate sufficiently high proportions of exchanges.

5.5.2 Re-evaluation of model behaviour

In our re-evaluation of model behaviour, comparing model performance at many parameter settings to the newly determined benchmarks, whilst observing the limits on error rate and non-contextuality determined in the previous chapter, we first considered generation of anticipations and perseverations, and then finally looked at exchange error generation.

Our results showed that the model was capable of generating appropriate proportions of anticipations and perseverations for the incompletes-as-anticipations bounds, but not the proportional-incompletes bounds when models which failed the constraints on error rate and non-contextuality of errors were excluded. However, the number of parameter settings at which the model met these bounds was exceedingly small. Proportion calculations are affected by errors generated in every category however, and we noted that it would be difficult to meet these bounds if the number of exchange errors being generated was very low, and we indeed later showed that this was the case.

However, before considering exchange error generation, some interesting anticipation and perseveration results independent of exchange error generation were uncovered. While the model tended to generate either too many anticipations and too few perseverations, or too many perseverations and too few anticipations, the specific models which generated too many perseverations were mostly excluded when the constraints on error rate and non-contextuality of errors were applied. This fits in with the result reported in the previous chapter that second onset errors produced by the model are frequently non-contextual. A higher rate of second onset contextual errors is therefore accompanied by an increase in non-contextual errors, and therefore an increase in error rate overall.

We linked this finding to Dell, Burger, and Svec's (1997) empirical result that the proportion of anticipation errors generated by a speaker, given the number of anticipations and perseverations generated by that speaker overall, is negatively correlated with overall error rate. The same negative correlation was shown to exist across our specific models. By varying spreading activation parameters therefore, the spreading activation model can account for this variance in human speakers. Specifically, manipulations of connection strength, the number of steps before selection, and decay rate cause this negative correlation, whereas manipulations of jolt to prime ratio and levels of activation-based noise don't, providing some alignment between the behaviour of the current model and the behaviour of the abstract

mathematical model presented by Dell, Burger, and Svec (1997). It was additionally shown that the spreading activation model predicts that as the proportion of anticipatory errors increases, the proportion of non-contextual errors should also decrease. This finding is in line with Schwartz et al.'s (1994) suggestion that "good" errors are more likely when speakers generate high proportions of anticipations, providing a prediction that can be empirically tested either by considering populations in which error rate is naturally elevated, or by manipulating error rate experimentally.

However, the most important result of this investigation was the demonstration that the model is incapable of generating a sufficiently high proportion of exchange errors without breaking other bounds imposed on its behaviour. Specifically, a small number of models generated proportions of exchange errors which were high enough for both analyses, but once the constraints on error rate and non-contextuality were applied, all models which generated more than the 14.3% exchanges required for the proportional-incompletes bounds were eliminated, and only two specific models which generated more than the 5.0% exchanges required by the incompletes-as-anticipations bounds remained. These final two models generated too many anticipations and not enough perseverations however.

We suggested that the triggering mechanism in the model is not strong enough, such that the activation which remains on an unselected first onset is too weak. Models which generate sufficiently high enough exchange proportions are instead relying on inappropriately high error rates, which increase the chance of coincidental first and second onset errors. Our explorations of how the parameter manipulations affected exchange error generation highlighted again that more exchanges are generated at low jolt to prime ratios, and more perseverations at higher jolt to prime ratios, helping to explain why the most successful exchange error generating simulations exhibit so few perseverations. It was also suggested that increasing the connection strength and decreasing the decay rate past values used in the literature (with the exception of one outlier decay rate setting in Dell, 1990) may help increase second onset error rate, although then measures such as reduction of activation based noise would probably be necessary to reduce the overall error rate. However, very limited further manipulation of these variables is possible as Shrager et al. (1987) have shown that connection strength must be lower than decay rate or activation levels in a network will rise without bound. It is therefore not clear that these manipulations would be sufficient to close the gap between the empirical benchmarks and the current behaviour of the model.

Even if further investigation can uncover parameter settings at which exchange error proportions rise substantially, this study has clearly demonstrated that across a very large portion of parameter space, covering nearly all parameter settings used in previous investigations of normal speech in Dell's (1986) model, the model as it stands is far from able to account for this corpus exchange error evidence.

5.5.3 Outlook

Here we consider what repercussions these results have for future work and for continuation of the work presented in this thesis.

Steps for the future

The current results show that evaluations of empirical speech error corpus results and evaluations of the spreading activation model's behaviour do not correspond. One approach to solving this problem would be to suggest that the data is simply not reliable. Speech error collection is clearly very open to collector bias, and our analysis highlights that there are huge differences in patterns across different corpora. Having said this, we allowed for massive variation in our benchmarks, and the model still was not able to generate enough exchanges. The only way to remedy this problem with empirical data would therefore be to demonstrate that humans generate hardly any exchanges at all.

If we however assume that 5% is a liberal lower limit on the proportion of exchange errors generated by normal speakers, then we find that there are no models in the literature which can account for this evidence. As noted in chapter 2, very few models address the problem of exchange error generation. The one other implemented model which successfully generates exchanges is presented by Vousden et al. (2000). Their results show that the model exhibits an exchange error proportion of 8.1%, calculated as a proportion of all anticipations, perseverations and exchanges, which is above the incompletes-as-anticipations lower exchange error bound, although still below the proportional-incompletes bound. However, this model also demonstrates a prohibitively high error rate of 15.6%, much higher than the 5.75% upper limit on error rate established in the previous chapter.

Attempts to fix the current model would in an ideal world first focus on the model's inability to generate incomplete errors. Incomplete errors form a very substantial portion of speech error corpora, which makes the current standard approach of trying to model them as complete errors somewhat unsatisfactory. To produce such

errors however, the model would require an implementation of a monitor-editor, and such implementations are currently rare even in models which rely on them as a core theoretical tenet (e.g., Levelt et al., 1999).

An approach more rooted in the current implementation would be to consider a prolonged period of post selection activation suppression. Instead of simply setting the selected phoneme's activation level to zero, phoneme activation could be repeatedly set to a fraction of its real value for a number of steps following selection. This would increase the trigger strength of the anticipation, as it would also reduce the second onset jolt activation passed to the anticipated onset. Simulations would be required to establish whether this would impact the number of perseverations generated too severely however.

More generally however, this work shows that the explanation of these errors which has stood for the past two decades is currently not able to account for this basic speech error evidence, highlighting a significant opening for future theoretical development.

Steps for the present

Our simulations demonstrate that the spreading activation model clearly has problems with contextual error generation on the second onset. The previous chapter showed that the model tends to generate an extremely high proportion of non-contextual errors on the second onset, and the current chapter uncovered a particularly extreme problem with exchange error generation.

The instrumental data we wish to focus on for the rest of this thesis all directly concerns the influence of competing onsets. It is therefore important that contextual error generation is operating correctly, such that the onset in question is a sufficiently strong competitor compared to other onsets. For the rest of this thesis, we therefore restrict evaluation of the model behaviour to behaviour on the first onset, where another onset is directly primed and do not consider anticipation, perseveration and exchange behaviour further. As problems with word sequencing reflect the implementation of the frame-and-slot model, they should not necessarily affect the model's ability to capture patterns assumed to reflect the interaction of processes, such as the lexical bias and phonological similarity effect, as these are explained by the implementation of spreading activation. We further note that the separation between these two aspects of the model may leave some scope for later resolution

of these movement error problems, such that improvements in the spreading activation component of the model could potentially be later reconnected with successful movement error simulation.

5.6 Chapter summary

In this chapter, we established new benchmarks for human anticipation, perseveration and exchange error generation, using multiple corpora, and applying two different approaches to interpreting incomplete errors. We applied these new benchmarks when investigating the spreading activation model's behaviour at multiple parameters, taking the limits on error rate and non-contextuality as determined in the previous chapter into account. It was shown that by manipulating parameters, the spreading activation model can account for a negative correlation between the proportion of anticipations generated and error rate, as empirically demonstrated by Dell, Burger, and Svec (1997). The simulations further predicted that a high proportion of anticipations should be associated with a low proportion of non-contextual errors, a prediction suitable for empirical verification.

Crucially, these investigations demonstrated that the spreading activation model cannot generate enough exchange errors to account for speech error corpus evidence, without generating too many errors (where a reduction in correct productions naturally leads to an increase in exchanges), or too few perseverations. Whilst limited possibilities exist for further parameter settings to be tested, the very large parameter space examined in this study ranges across nearly all the parameter space tested in the literature to date. It was argued that the trigger mechanism in Dell's (1986) model does not appear to be strong enough. To explain this discrepancy between the model behaviour and the corpus evidence as purely a collector bias effect on the corpora, it would be necessary to obtain data demonstrating that humans in fact produce an extremely small proportion of exchange errors.

For the current thesis, we do not consider productions on the second onset any further, and instead focus on investigating interactions between levels of representations and their impact on first onset error patterns.

Chapter 6

Statistical methods for large scale modelling: with classic results as test cases

6.1 Introduction

Many results which are used to make claims about human word production take the form of statistical evidence that behaviour differs between given conditions. For example, in an experimental situation where half the materials were set up so that onset errors would result in words, and half were set up so that onset errors would result in non-words, a greater number of errors would be expected in the word outcome condition, a result known as the lexical bias effect (e.g. Hartsuiker et al., 2005). Within Dell's (1986) model, the higher number of lexical outcome errors is used to argue for feedback from phonemes to words.

The instrumental evidence which we wish to model in order to determine constraints on activation flow between phonemes and features in a model with output at the subphonemic level also takes this form. For example, Goldrick and Blumstein (2006) showed that where an intended voiceless consonant (e.g., /k/) was produced as a voiced consonant (e.g., [g]), the resulting voiced consonant was more voiceless than an intended and correctly produced voiced consonant. This result was used to argue that activation from the intended but unselected voiceless consonant must cascade to subphonemic representations. Similarly, McMillan (2008) showed that articulations of a phoneme given a competing phoneme which differed by a single feature were less like a reference measurement of the articulation of the target phoneme than in a condition when the competing phoneme differed by two features. This result was used to argue that activation from subphonemic representations must feed back to phonemes.

In modelling such results, we argue that it is not sufficient to show that the model exhibits a numerical difference between conditions. In all the specific models which we consider here, random noise affects the activation levels of the nodes, as is the case for most implementations of Dell's (1986) model (including Dell & Gordon, 2003; Dell et al., 2004; Dell, Schwartz, et al., 1997; Hartsuiker, 2002; Foygel & Dell, 2000; Goldrick & Rapp, 2002; Martin et al., 1994; Oppenheim & Dell, 2008; Rapp & Goldrick, 2000; Ruml & Caramazza, 2000; Ruml et al., 2000, 2005; Schwartz et al., 2006). It is therefore important to check that any differences that the model exhibits between conditions are statistically reliable, and cannot be explained by chance.

We have argued however that it is not appropriate to test different architectures at just one parameter setting. Results may be accounted for in different ways in different architectures. For example, Goldrick and Blumstein (2006) claimed that in a model with output at a subphonemic level, cascading from all phonemes was required to account for their results. However, in chapter 2, we proposed two mechanisms by which models with no cascading from phonemes and models with cascading from selected phonemes only would also be able to explain these findings. We cannot rule out the possibility that the optimal parameter settings for the different mechanisms proposed may differ, and a core part of the approach taken in this thesis is therefore to test the behaviour of architectures at multiple parameter settings. This means that statistical tests of behaviour are run in 5832 different specific models for architectures with feedback, and 2916 different specific models for architectures without feedback.

The aim of this chapter is to address the question of how to evaluate whether a given architecture can really account for the findings, given that some of the significant results returned by statistical tests of model behaviour in an architecture may be due to Type I errors. We use the classic lexical bias and phonological similarity effects as a test case. Dell (1986) accounted for the lexical bias effect by positing feedback from phonemes to words, and similarly explained the phonological similarity effect by assuming feedback from features to phonemes. We note that Rapp and Goldrick (2000) have provided evidence that Dell's (1986) model generates a statistically significant lexical bias effect and that feedback from phonemes to words is required for this behaviour. However, the only simulation evidence in the literature that the model exhibits the phonological similarity effect is a report from Dell (1986) that errors share more features with the intended target as the number of steps before selection increases. This result is in line with the argument that feedback

from features to phonemes activates similar phonemes, as a greater number of steps before selection would allow activation to flow around these feedback loops for longer. Here, we provide statistical evidence that the model can account for both the lexical bias and phonological similarity effects, and confirm that feedback from phonemes to words is required for the model to exhibit a lexical bias, and feedback from features to phonemes is required for a phonological similarity effect. This allows us to demonstrate our methodology for assessing whether architectures can account for single effects, and whether they can account for multiple effects without requiring different parameter settings for different effects. We also show that these effects can be accounted for by specific models which respect the constraints on error rate and non-contextuality of errors which we derived from human data in chapter 4. Finally, we demonstrate how the parameter exploration methodology developed in chapter 4 can be extended to uncover which parameter settings lead models to demonstrate the significant effects of interest, and consider which of these parameter settings allow the error rate and non-contextuality constraints to be observed.

6.2 Simulation methodology

We ran simulations to determine which architectures can generate lexical bias and phonological similarity effects, and which parameter settings this requires. These used new materials which we outline in this section. To find out which specific models generated appropriate error rates and proportions of non-contextual errors given the constraints introduced in chapter 4, we also ran further random word generation simulations, as it was felt that behaviour on these simulations would be more representative of normal speech than behaviour on simulations with highly manipulated materials. This section provides further details of all of these simulations.

6.2.1 Model configuration

We consider four one-stage models of phonological encoding, by orthogonally varying the presence of phoneme-to-word feedback and feature-to-phoneme feedback. Parameter settings were varied for all models as outlined in section 3.6. As explained in section 3.6, architectures which contain no feedback from phonemes to words and no feedback from features to phonemes, the strength of feedback connectivity *fbkConn* is not varied as this would have no effect. Bar charts which compare the numbers of specific models falling into given behaviour categories across architectures therefore report percentages of specific models instead of raw counts, for

ease of comparability. Combining architectures with parameter settings, we tested 20,412 specific one-stage models in total.

6.2.2 Model task and lexicon

For the one-stage architecture with feedback from phonemes to words and from features to phonemes, we took error rate and non-contextuality data from the simulation described in chapters 4 and 5. To collect error rate and non-contextuality data for all other architectures, we ran simulations identical to the previous simulation with exactly the same list of random words. Due to the problems identified in those chapters with generation of contextual errors on the second word, we focused exclusively on productions of the first word for all simulations. Priming of the second word would obviously influence first word productions however. Similarly, simulations of the transcribed lexical bias and phonological similarity effect focused on single word productions, but for all productions a competitor was primed in the same way that an upcoming word was primed in our word pair production simulations. In both of these cases, the nature of the competitor can either be seen as a word from later in the phrase, or a word which was activated due to other higher level influences, such as semantic similarity or presence of a related object in the speaker's environment. Other modellers have investigated incidental errors in single word production with no primed competitor (e.g. Dell, Schwartz, et al., 1997; Rapp & Goldrick, 2000). However, we chose to use a primed competitor to keep the current simulations in line with studies in later chapters, which aim to explicitly simulate the influence of contextual competing phonemes as investigated in the experiments reported by Goldrick and Blumstein (2006), McMillan et al. (2009) and McMillan (2008).

The competitor was manipulated to create conditions in which production of the competing onset would result in a word, and conditions where it would not; and conditions where the competing phoneme was phonologically similar (sharing two features), and conditions where it was not (sharing one feature only). Lexicality of outcome and similarity of onset phonemes were crossed in this material set, with an intention to later examine possible interactions between these two variables. However, this analysis is not carried out in the current thesis. We note that a previous simulation using one set of parameters suggests that the spreading activation model numerically predicts no interaction (Oppenheim & Dell, 2008), although no statistical analysis of the model's results was reported.

Table 6.1: Materials for lexical bias and phonological similarity simulations, where place of articulation always differs between target and competitor onset.

	Differing onset features	
	Only place	Place and voicing
Lexical outcome	/k/-/t/ call torn /g/-/d/ gaud doors	/k/-/d/ cord dawn /g/-/t/ gall tours
Nonlexical outcome	/k/-/t/ cord tours /g/-/d/ gall dawn	/k/-/d/ call doors /g/-/t/ gaud torn

Table 6.2: Materials for lexical bias and phonological similarity simulations, where voicing always differs between target and competitor onset.

	Differing onset features	
	Only voicing	Place and voicing
Lexical outcome	/k/-/g/ call gaud /t/-/d/ torn doors	/k/-/d/ corn doored /g/-/t/ gauze tall
Nonlexical outcome	/k/-/g/ corn gauze /t/-/d/ tall doored	/k/-/d/ call doors /g/-/t/ gaud torn

The materials designed here were also intended for use in simulations of Goldrick and Blumstein's (2006), McMillan et al.'s (2009) and McMillan's (2008) articulatory lexical bias and phonological similarity evidence. We therefore used the velar and alveolar stop onsets /k/, /g/, /t/ and /d/ which were used in these previous studies. Two material sets were created: one in which place always differed between target and competitor, for use in simulating EPG and ultrasound studies where the influence of a competitor with a different place of articulation is investigated; and one in which voicing always differed between target and competitor, for use in simulating studies measuring VOT, where the influence of a competitor with different voicing is investigated. Behaviour of the specific models was evaluated separately for the two different sets of materials.

The full material sets are shown in table 6.1 and table 6.2. Pairs were tested both with the first word listed as the target word and the second word listed as the competitor, and vice versa. For example, where *"call torn"* is listed as a pair, a target *"call"* with primed competitor *"torn"* was tested, as well as a target *"torn"* with primed competitor *"call"*.

To create these material sets, we chose one nucleus vowel and four coda consonants, such that each coda, when combined with the nucleus vowel, produced a word for three of our target onsets, and a non-word for the remaining onset (or alternatively, an English word which was left our of our model's lexicon and which was therefore deemed a non-word for the current analysis). Using these codas and the chosen nucleus, it was therefore possible to create a word outcome pair (where both the target pair and the outcome pair were words) and a non-word outcome pair for each possible pair of onsets.

A number of variables were controlled for in our design, many of which concerned frequency of components, or how many times a node is connected to nodes at a higher level. In Dell's (1986) model, the amount of activation sent from one node to another depends solely on the amount of activation the sender node has and the strength of the connection from the sender node to the receiver node. Activation transmitted from the sender node is not "shared" between receiver nodes; i.e., it does not decrease as the number of receiver nodes increases. Especially when feedback connections are present, this behaviour leads to frequent nodes becoming more activated, as more nodes reflect back the activation of the frequent node. Frequency of co-occurrence of representations is also important, as representations which frequently occur together will boost each other's activation levels via shared higher level nodes. We therefore constructed the material sets such that all onsets, all codas, and the chosen nucleus were used in every condition. This prevented the frequency of onsets, nuclei and coda phonemes and features from differing across conditions. Similarly, the same onset-nucleus combinations and nucleus-coda combinations were used in each condition, thereby controlling for the frequency of their occurrence too. To control onset-coda frequency, we created a lexicon in which there were no other words using combinations of our chosen onsets and codas, as it was otherwise too difficult to find appropriate vocabulary to keep the number of words for each combination constant. Similarity of target items and components was also a concern. In the same way that similar competing onsets would be expected to receive more activation from target onsets than dissimilar competing onsets, due to the higher number of features shared, similar nuclei and coda would also become more activated, which through feedback loops would also affect the activation of the word they participated in and the onset phoneme. Choosing one nucleus, and voiced alveolar coda consonants which differed only in manner features, made it possible to avoid these similarity problems.

Each of the resulting 16 target and competitor combinations in each material set was produced 500 times by each specific model, resulting in a total of 8000 word productions.

The full lexicon contained 100 CVC words: the 12 target words, and 88 others chosen from the BEEP lexicon (Robinson, n.d.). A larger lexicon was used than in the previous simulations as so many target words were chosen and these bore great similarity to each other. Only words with one known pronunciation were allowed, and vulgar words were not permitted. Furthermore, no words where the vowel was a diphthong were included, to simplify vowel representation in the network. As noted above, the words *"cause"* and *"toured"* were also excluded from the model's lexicon and no words using combinations of the experimental onsets and codas were permitted. The lexicon is presented in table 6.3.

We tested models using both the set of materials in which the place always differs between target and competitor onset, and the set of materials in which the voicing feature always differs between target and competitor onset. This was done because an understanding of the model's behaviour with respect to the categorical transcribed lexical bias and phonological similarity could provide a useful reference when evaluating the model's attempts to capture the VOT, EPG and ultrasound lexical bias and phonological similarity evidence in later chapters. For simplicity, in this chapter we present results from the simulations using the material set in which place always differs between target and competitor onset, but analyses verified that results from simulations using the other material set did not significantly differ.

6.2.3 Onset error classification

Output was determined based on the selected phonemes, as in the previous chapters. Onset productions were then further classified as *correct productions*, if the intended onset was produced; *contextual errors*, if the competing onset was produced; and *non-contextual errors* if any other onset was produced.

In the lexical bias and phonological similarity simulations, statistical analyses focused on the number of contextual errors produced in the different conditions. For each specific model, a logistic regression was used to determine whether lexicality had a significant effect on the number of contextual errors produced in the predicted direction. Specific models in which the logistic regression demonstrated that there were more contextual errors than would be predicted by chance (given $\alpha = 0.05$) were deemed to show a lexical bias. In these first investigations, we used a one-tailed

Table 6.3: The lexicon of the model for the current simulation.

Target words			
call	dawn	gall	tall
cord	doored	gaud	torn
corn	doors	gauze	tours

Other words			
bang	harsh	mousse	shack
bat	hawk	muck	shard
bawl	hen	niece	sheen
bin	hitch	nun	soon
booze	hoot	pal	soothe
buck	jim	pang	suit
bud	josh	pawn	sum
bug	june	peal	thatch
chalk	keep	peep	this
chap	knot	peg	thud
cheese	lad	pet	thug
chin	lass	phil	tom
chute	leap	piece	toot
coot	lees	pub	vet
course	let	rev	wash
ditch	loch	rig	whiff
fetch	loot	rim	whim
fit	lose	root	wick
fool	loss	rot	win
gawp	mall	rub	wreath
give	marge	rude	zen
hap	match	rug	zoom

test as the outcome of interest was solely whether models demonstrated the human behaviour pattern. The fact that a one-tailed test was used is accounted for in our analysis of whether an architecture can account for an effect. Similarly, a logistic regression was used to determine whether phonological similarity had a significant effect on the number of contextual errors produced in the predicted direction. As noted in section 6.2.2, in this thesis we do not report on an analysis crossing lexicality and phonological similarity and determining whether an interaction is predicted by the model.

6.3 Determining which architectures can account for the lexical bias and phonological similarity effects

In this section, we introduce our methodology for establishing whether an architecture can account for an effect when statistical tests are carried out at many different parameter settings, using the well established transcribed lexical bias and phonological similarity effects as test cases. We expect to show that feedback from phonemes to words is required to account for the lexical bias effect, and feedback from features to phonemes is necessary to account for the phonological similarity effect. We show that our methodology can be extended to verify that an architecture can account for multiple effects without requiring different parameter settings for the separate effects. Using this approach, we expect to demonstrate that there are specific models with feedback from phonemes to words and from features to phonemes which can account for both the lexical bias and the phonological similarity effect.

To begin, we examined the differences in error rate and proportions of non-contextual errors generated by the four different architectures, to help develop an understanding of their basic behaviour.

6.3.1 Error rate and non-contextuality of errors

In line with our findings in chapter 4 that an increase in feedback connection strength generally results in an increase in error rate and the proportion of non-contextual errors generated, figures 6.1 and 6.2 demonstrate higher error rates and non-contextuality proportions when either phoneme-to-word or feature-to-phoneme feedback is added to the model. Feature-to-phoneme feedback causes more of an increase in non-contextual error generation than phoneme-to-word feedback, suggesting that activation passing from target or primed features to other similar phonemes is a particular cause of non-contextual errors. Figure 6.3 reflects these two patterns.

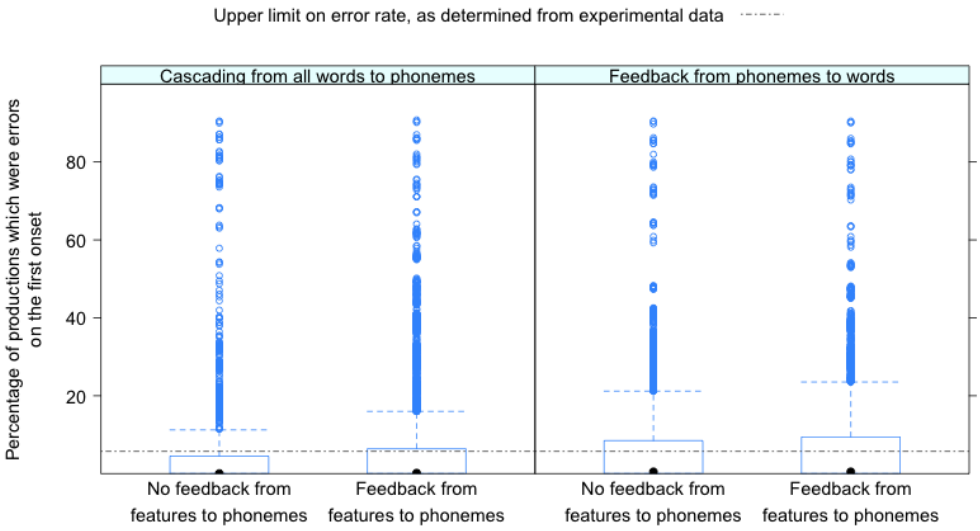


Figure 6.1: The effect of phoneme-to-word and feature-to-phoneme feedback on first onset error rate in one-stage models of phonological encoding. The dotted line represents the upper limit on error rate as calculated in chapter 4.

As feedback is added, the tendency for models to generate errors for analysis increases, but the tendency for models to fail the error rate or non-contextuality constraints increases too.

6.3.2 Activation flow options required to account for the lexical bias and phonological similarity effects

In this section, we introduce our method for determining whether a certain number of specific models showing significant results constitutes evidence that an architecture is capable of accounting for a given result or multiple results, using the lexical bias and phonological similarity effects as a case study.

Introduction of the binomial method: Lexical bias

Figure 6.4 clearly suggests that feedback from phonemes to words is required for a lexical bias effect to be exhibited. However, the graph also shows that some specific models with feedback from phonemes to words do not exhibit a significant lexical bias according to the logistic regression analysis carried out per model (as explained in section 6.2.3), whilst some models without feedback from phonemes to words do. Given that we have carried out such a high number of statistical tests, how do we determine whether the number of specific models showing significant results for a given architecture is sufficient evidence that an architecture can account for a

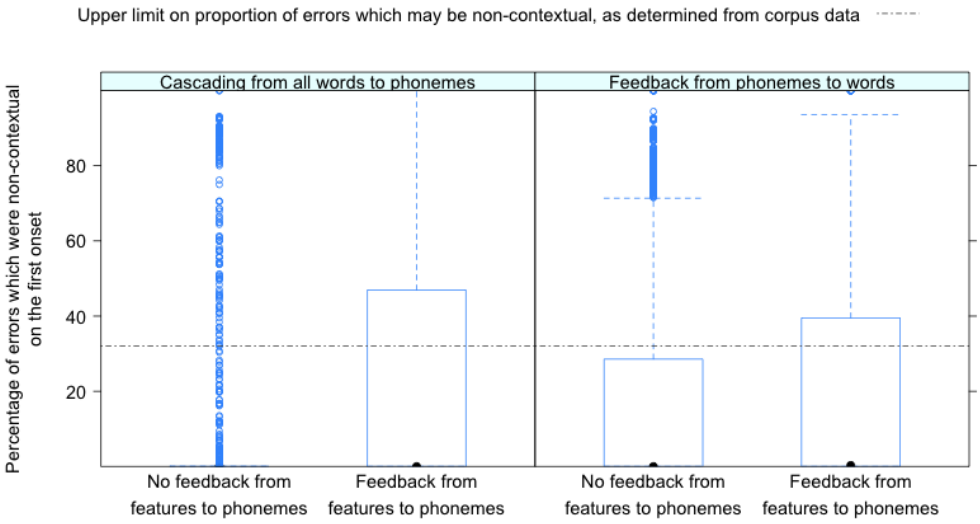


Figure 6.2: The effect of phoneme-to-word and feature-to-phoneme feedback on the proportion of errors which are non-contextual at the first onset in one-stage models of phonological encoding. This proportion can only be calculated for specific models which generated at least one error. The dotted line represents the upper limit on error non-contextuality as calculated in chapter 4.

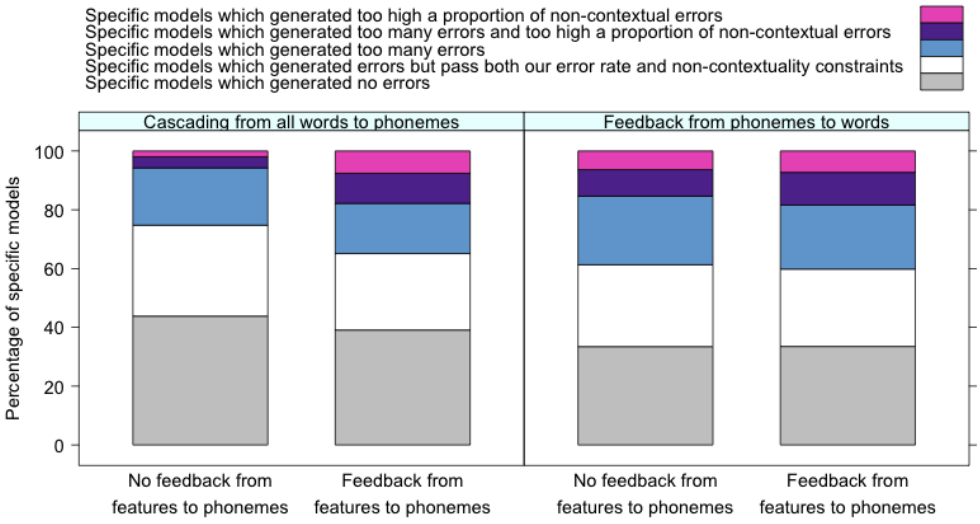


Figure 6.3: The effect of phoneme-to-word and feature-to-phoneme feedback on the numbers of specific models which pass our constraints on error rate and error non-contextuality, for all one-stage models. See figure 4.9 in chapter 4 for further elaboration on the key.

certain effect, or whether this number of significant results could be expected as a result of Type I errors?

For each individual test on each specific model, we set the probability of a Type I error to 0.05. We can use this knowledge to build a binomial model of how many Type I errors we would expect to occur for a given number of statistical tests. We can then use the binomial model to determine whether the number of specific models which returned significant results is likely to be due to chance. If calculations with the binomial model suggest that there is less than 0.05 chance of this number or greater of Type I errors occurring, then we accept these results as evidence that the architecture can account for the evidence. Otherwise we do not reject the null hypothesis that the findings can be explained as being due to Type I errors, or in other words, chance.

This test is run for each architecture separately. As we are testing four architectures, we apply the Bonferroni correction, and only rule that there is sufficient evidence of an architecture being able to account for the human result when a large enough number of specific models display significant results for the binomial model to demonstrate that there would be less than 0.0125 (0.05/4) probability of observing this number or greater significant results due to Type I errors. Models which generated at least one contextual error are considered as the models for which statistical tests were run, as it is clear that models which generate no contextual errors will not display a significant effect as a result of Type I error or otherwise.

However, these calculations paid no regard to whether specific models in consideration failed the constraints on error rate and non-contextuality as determined in chapter 4. We therefore re-evaluated the behaviour of the different architectures when specific models which failed the constraints were excluded. We note that excluding these models is likely to increase the proportion of tested models which have too little power for a significant effect to be displayed due to too few errors being generated. However, it would be useful to demonstrate that specific models exist that both pass the human data derived constraints on error rate and non-contextuality of errors, and display lexical bias effects as humans have been shown to.

Figure 6.5 suggests that exclusion of constraint failing models does not change our overall conclusions, such that there is still a much larger number of models displaying significant lexical bias effects for architectures with feedback from phonemes to words than there is for architectures without this feedback. Table 6.5 shows the

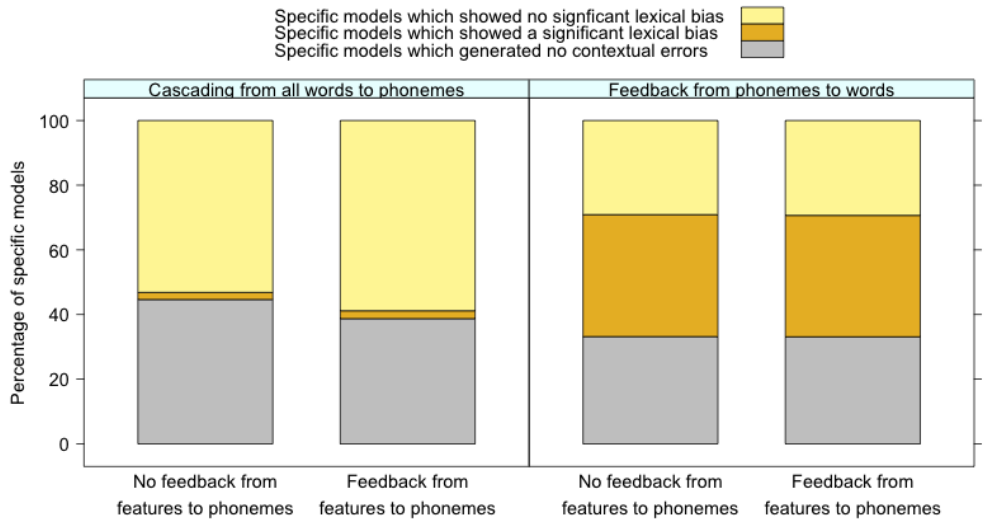


Figure 6.4: The effect of phoneme-to-word and feature-to-phoneme feedback on exhibition of lexical bias effects in one-stage phonological encoding models.

Table 6.4: Binomial analysis to determine which one-stage architectures can generate a lexical bias effect. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.0125) of witnessing the same number or more specific models generating significant lexical bias effects by chance.

	Specific model counts			**Prob.**	
	Total	Generated contextual errors	Significant lexical bias		
Cascading from all Ws to Ps					
No feedback from Fs to Ps	2916	1614	64	$> .9$	
Feedback from Fs to Ps	5832	3573	142	$> .9$	
Feedback from Ps to Ws					
No feedback from Fs to Ps	5832	3897	2200	$< .001$	*
Feedback from Fs to Ps	5832	3901	2190	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

results of a binomial analysis where specific models which fail the constraints are excluded, and we only consider the remaining models which generated contextual errors as tested. Again, our calculations show that there is plenty of evidence that architectures with feedback from phonemes to words can account for this effect, but insufficient evidence that models with no feedback from phonemes to words can capture this behaviour pattern.

Further demonstration of the use of the binomial method: Phonological similarity

We then applied the same approach to investigating which architectures could exhibit the phonological similarity effect. Figure 6.6 clearly suggests that feedback from features to phonemes is required for a phonological similarity effect to be exhibited. A binomial analysis of the number of significant results found for each architecture, as summarised in table 6.6, leads to the same conclusions.

Again, excluding models which fail the constraints on error rate and non-contextuality of errors does not change this conclusion, as demonstrated in figure 6.7 and by the binomial analysis summarised in table 6.7.

The binomial method and multiple effects: Lexical bias and phonological similarity

Finally, we show that the binomial method can also be used to investigate which architectures can account for multiple effects simultaneously; i.e., without requiring different parameter settings for the different effects. However, we argue that our current approach results in a reduction of power as more effects must be accounted for. We use the lexical bias effect and the phonological similarity effects as case studies.

When investigating the ability of architectures to account for single effects, we noted that the probability of a Type I error on one specific model was 0.05. When accounting for multiple effects, the probability that at least one of the significant effects is due to chance increases. For example, for two effects, the probability that at least one of the significant effects is due to chance is $1 - (0.95 * 0.95) = 0.0975$. A higher chance of a Type I error on each specific model means that more specific models must demonstrate significant effects (here, for both results) in order for us to conclude that there is evidence that the architecture can account for the result. Given that the number of specific models which show significant effects for two effects can never be higher than the number of specific models which show significant effects for one of these effects, this must mean that binomial analyses of

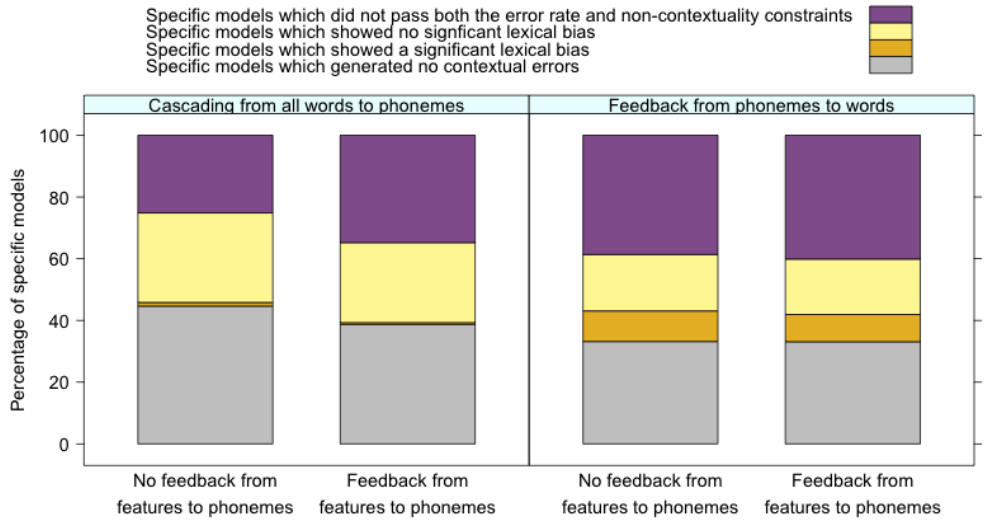


Figure 6.5: The effect of phoneme-to-word and feature-to-phoneme feedback on exhibition of lexical bias effects in one-stage phonological encoding models, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

Table 6.5: Binomial analysis to determine which one-stage architectures can generate a lexical bias effect, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.0125) of witnessing the same number or more specific models generating significant lexical bias effects by chance.

	Specific model counts				**Prob.**	
	Total	Excluded	Generated contextual errors	Significant lexical bias		
Cascading from all Ws to Ps						
No feedback from Fs to Ps	2916	734	880	35	$> .9$	
Feedback from Fs to Ps	5832	2028	1545	33	$> .9$	
Feedback from Ps to Ws						
No feedback from Fs to Ps	5832	2253	1644	578	$< .001$	*
Feedback from Fs to Ps	5832	2339	1562	514	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

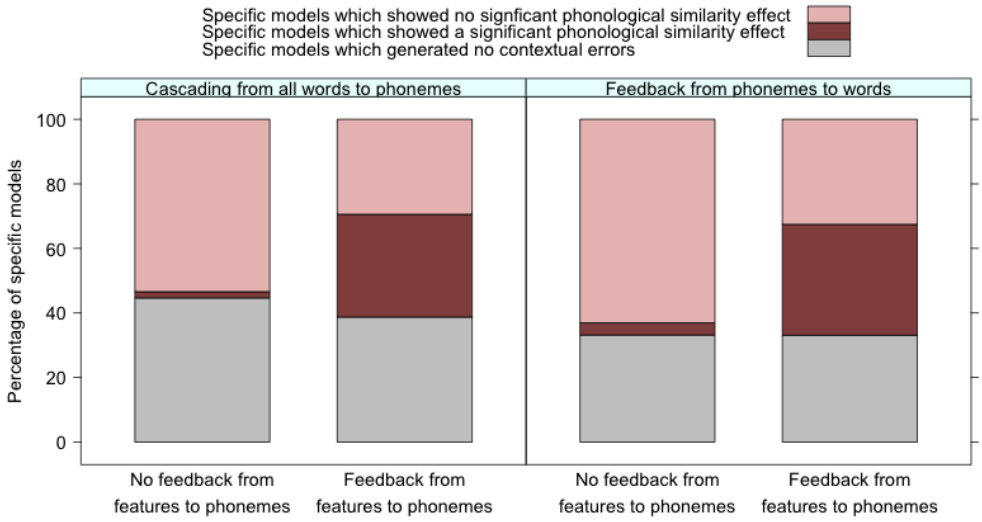


Figure 6.6: The effect of phoneme-to-word and feature-to-phoneme feedback on exhibition of phonological similarity effects in one-stage phonological encoding models.

Table 6.6: Binomial analysis to determine which one-stage architectures can generate a phonological similarity effect. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.0125) of witnessing the same number or more specific models generating significant phonological similarity effects by chance.

	Specific model counts			**Prob.**	
	Total	Generated contextual errors	Significant phonological similarity effect		
Cascading from all Ws to Ps					
No feedback from Fs to Ps	2916	1614	57	$> .9$	
Feedback from Fs to Ps	5832	3573	1857	$< .001$	*
Feedback from Ps to Ws					
No feedback from Fs to Ps	5832	3897	219	0.037	
Feedback from Fs to Ps	5832	3901	2003	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

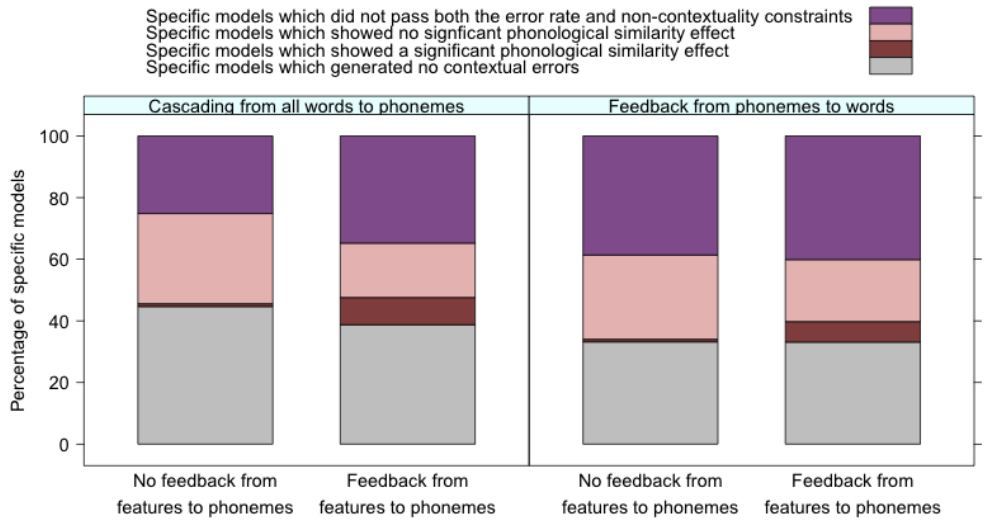


Figure 6.7: The effect of phoneme-to-word and feature-to-phoneme feedback on exhibition of phonological similarity effects in one-stage phonological encoding models, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

Table 6.7: Binomial analysis to determine which one-stage architectures can generate a phonological similarity effect, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.0125) of witnessing the same number or more specific models generating significant phonological similarity effects by chance.

	Specific model counts				**Prob.**	
	Total	Excluded	Generated contextual errors	Significant phonological similarity effect		
Cascading from all Ws to Ps						
No feedback from Fs to Ps	2916	734	880	27	$> .9$	
Feedback from Fs to Ps	5832	2028	1545	517	$< .001$	*
Feedback from Ps to Ws						
No feedback from Fs to Ps	5832	2253	1644	50	$> .9$	
Feedback from Fs to Ps	5832	2339	1562	389	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

whether an architecture can account for two effects has less power than a binomial analysis of whether an architecture can account for a single effect, and that power reduces further as more effects must be accounted for. An improvement to the method to remove this power reduction is an area for future research. However, using the current method we can rely on any positive results demonstrating that there is evidence that a model can account for multiple effects simultaneously.

As Dell's (1986) original claims and our previous results would predict, figure 6.8 shows that only the architecture with both feedback from phonemes to words and feedback from features to phonemes exhibits a large number of specific models for which both the lexical bias and phonological similarity effects are significant. The binomial calculation summarised in table 6.8 serves to confirm that the number of specific models exhibiting both significant lexical bias and significant phonological similarity effects is sufficient to conclude that the architecture with both feedback from phonemes to words and feedback from features to phonemes can simultaneously account for both of these effects. Again, the same conclusions are drawn from an analysis in which specific models which fail the constraints on error rate and non-contextuality of errors are excluded, as shown in figure 6.9 and by the binomial analysis summarised in table 6.9.

6.4 Exploring the parameter settings required for the lexical bias and phonological similarity effects

In the previous section, we demonstrated how our binomial method can be used to show that feedback from phonemes to words is required for models to exhibit the lexical bias effect, and feedback from features to phonemes is required for models to exhibit the phonological similarity effect. We extended this method to show that architectures with both sorts of feedback can account for both the lexical bias and phonological similarity effect simultaneously. However, not all specific models with feedback from phonemes to words show a lexical bias effect, and not all specific models with feedback from features to phonemes show a phonological similarity effect. Yet within the architecture with both types of feedback, there seems to be a big overlap between the set of specific models which display the lexical bias effect, and the set of specific models which display the phonological similarity effect.

In this section, we apply the methodology introduced in chapter 4 to gain some insight into which parameter settings lead models with feedback from features to

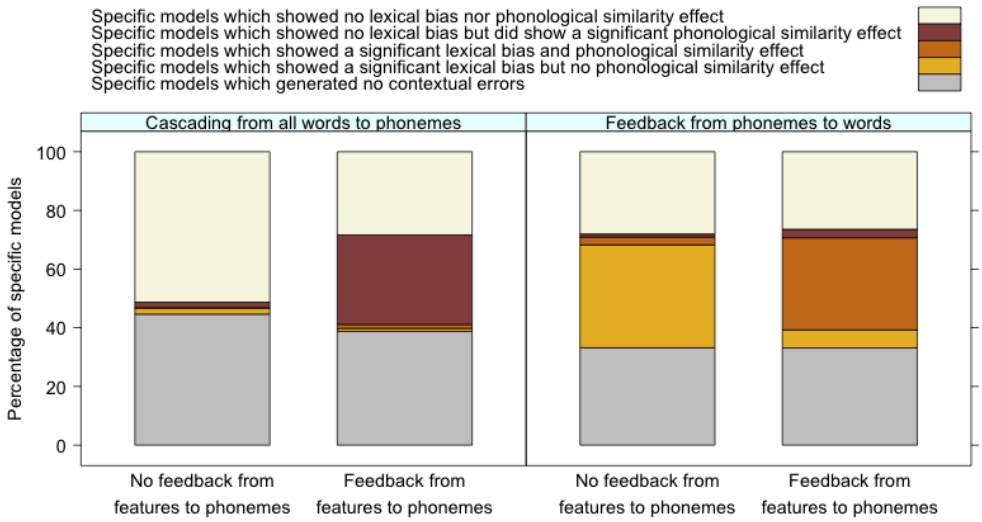


Figure 6.8: The effect of phoneme-to-word and feature-to-phoneme feedback on exhibition of lexical bias and phonological similarity effects in one-stage phonological encoding models.

Table 6.8: Binomial analysis to determine which one-stage architectures can generate both a lexical bias and a phonological similarity effect. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.0125) of witnessing the same number or more specific models generating both a significant lexical bias and a significant phonological similarity effect by chance.

	Specific model counts			**Prob.**	
	Total	Generated contextual errors	Significant LB and PS effects		
Cascading from all Ws to Ps					
No feedback from Fs to Ps	2916	1614	3	$> .9$	
Feedback from Fs to Ps	5832	3573	81	$> .9$	
Feedback from Ps to Ws					
No feedback from Fs to Ps	5832	3897	157	$> .9$	
Feedback from Fs to Ps	5832	3901	1833	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

LB = lexical bias, PS = phonological similarity

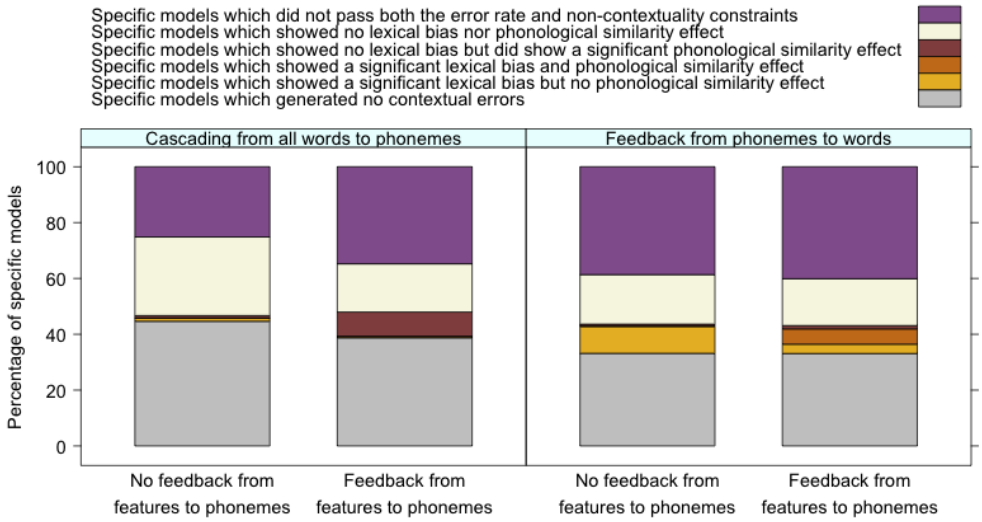


Figure 6.9: The effect of phoneme-to-word and feature-to-phoneme feedback on exhibition of lexical bias and phonological similarity effects in one-stage phonological encoding models, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

Table 6.9: Binomial analysis to determine which one-stage architectures can generate both a lexical bias and a phonological similarity effect, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.0125) of witnessing the same number or more specific models generating both a significant lexical bias and a significant phonological similarity effect by chance by chance.

	Specific model counts				**Prob.**	
	Total	Excluded	Generated contextual errors	Significant LB and PS effects		
Cascading from all Ws to Ps						
No feedback from Fs to Ps	2916	734	880	2	$> .9$	
Feedback from Fs to Ps	5832	2028	1545	13	$> .9$	
Feedback from Ps to Ws						
No feedback from Fs to Ps	5832	2253	1644	18	$> .9$	
Feedback from Fs to Ps	5832	2339	1562	319	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

LB = lexical bias, PS = phonological similarity

phonemes and feedback from phonemes to words to generate lexical bias and phonological similarity effects. This will help clarify why some models of this architecture generate these effects and others do not, illuminating in what way lexical bias and phonological similarity generation is dependent on the parameter settings used, and helping us understand why this overlap between models which generate lexical bias and models which generate phonological similarity exists. For logistic regression analyses of the effects of manipulating the spreading activation parameters, the dependent variable was whether specific models showed a significant effect. Each specific model therefore contributes 1 measurement to these logistic regressions, rather than 10,000 as in the previous chapters. As a result, effect sizes are smaller and some parameters do not display significant results.

Whilst previous chapters have already examined the behaviour of this architecture in great detail, in this chapter we focus on first word productions only. A graph of the effects of manipulating the spreading activation parameters on the number of specific models which pass or fail the two constraints on error rate and non-contextuality is therefore provided, as our assessment of which specific models passed or failed these constraints in figure 4.9 in chapter 4 was based on behaviour on both onsets. Figure 6.10 shows that the change of focus to the first onset only does not change the patterns greatly.

Figure 6.11 demonstrates that parameter settings which tend to result in lexical bias also tend to result in phonological similarity effect. Logistic regression analyses of the effects of manipulating the spreading activation parameters as summarised in tables 6.10, 6.11 and 6.12 tell a similar story. Note that only specific models which generated at least one contextual error were included in these regressions, to increase the extent to which the regression results reflect whether models generate a lexical bias or phonological similarity effect, rather than whether any data is available for this analysis. Specifically, high connectivity strengths, low jolt to prime ratios, high numbers of steps before selection, and high levels of activation-based noise all increase the numbers of specific models showing significant lexical bias and phonological similarity effects. No significant effect of manipulating the decay rate or the intrinsic noise parameter is found.

We suggest that there are two reasons that manipulating parameter settings affects the probability that a lexical bias or phonological similarity effect is found. Firstly, a higher error rate is likely to increase the power of the logistic regression analysis carried out on the behaviour of the specific model. Secondly, some parameters will increase activation flow through the feedback loops in the model which underlie

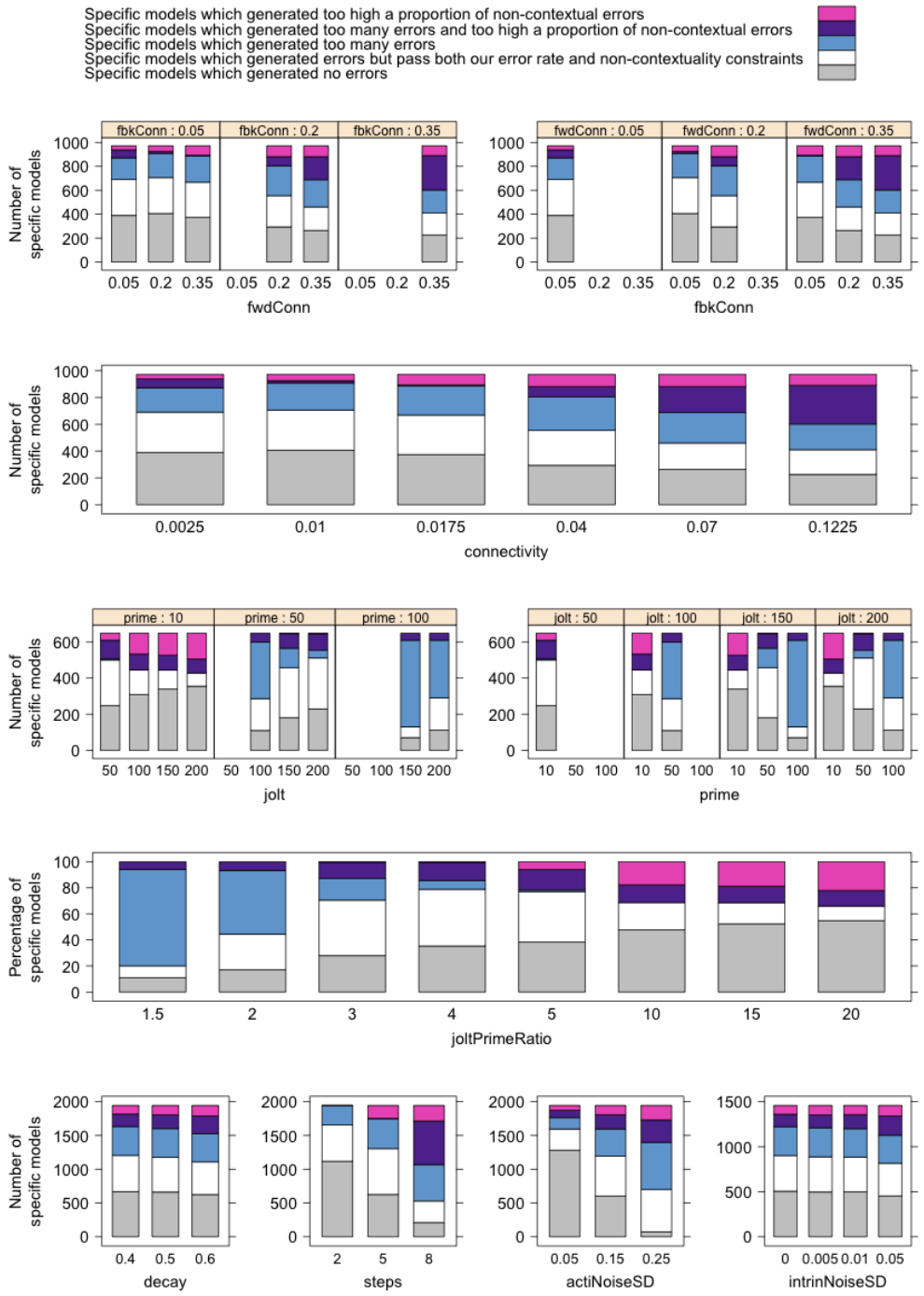


Figure 6.10: The effect of changing parameter settings on the numbers of specific models which pass our constraints, considering productions on the first word only, for specific one-stage models with feedback from phonemes to words and from features to phonemes. See figure 4.9 in chapter 4 for further elaboration on the key.

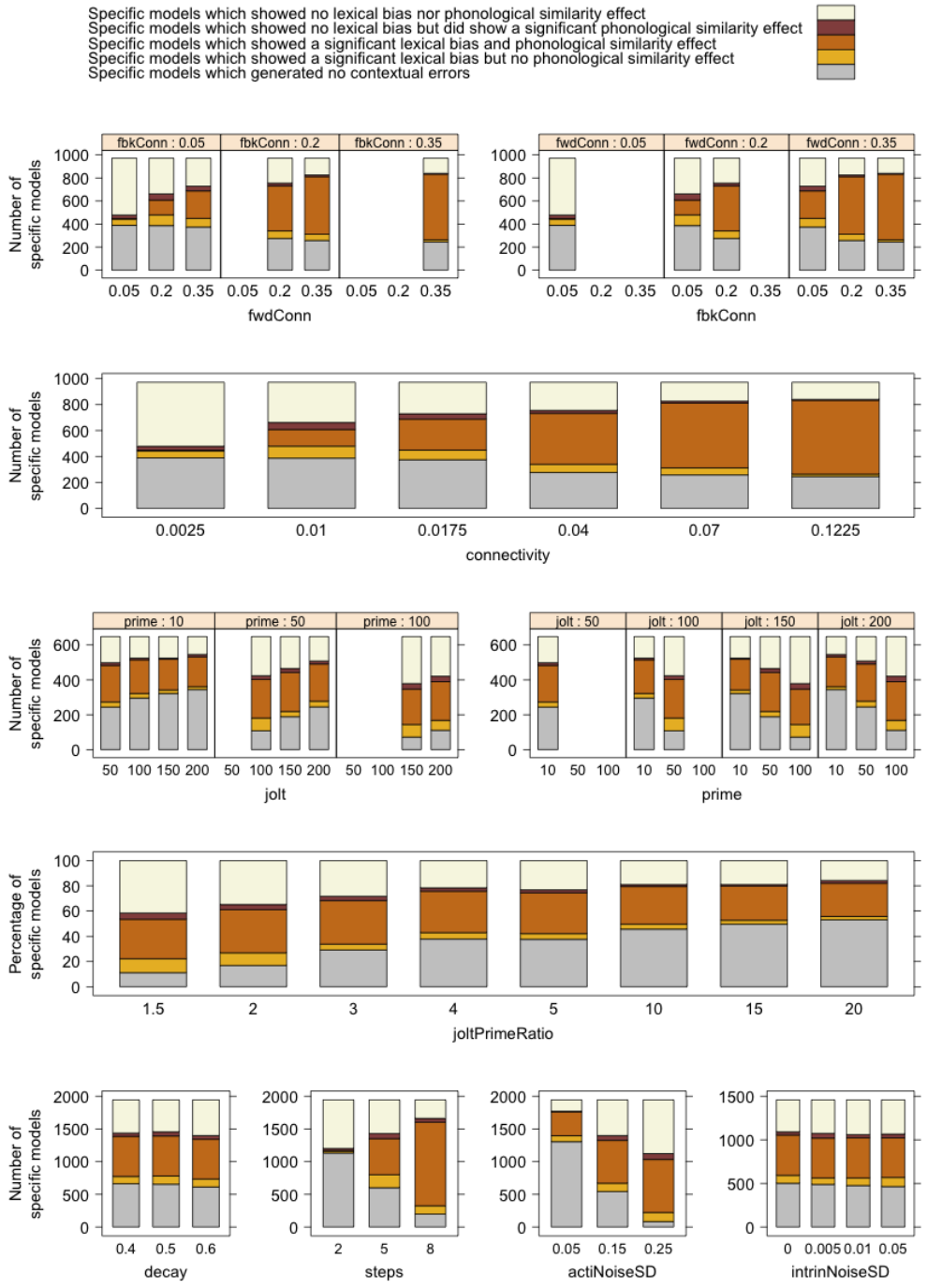


Figure 6.11: The effect of changing parameter settings on exhibition of lexical bias and phonological similarity effects in one-stage phonological encoding models with feature-to-phoneme feedback.

Table 6.10: Results of logistic regression model analyses using parameter values to predict the occurrence of lexical bias effects, for all one-stage models with phoneme-to-word and feature-to-phoneme feedback connectivity which generated at least one contextual error. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	+	25.5	1100	< .001	*
joltPrimeRatio	−	5.0	25	< .001	*
decay	−	0.8	1	0.434	
steps	+	29.9	1722	< .001	*
actiNoiseSD	+	8.6	78	< .001	*
intrinNoiseSD	+	0.1	0	0.885	

Table 6.11: Results of logistic regression model analyses using parameter values to predict the occurrence of phonological similarity effects, for all one-stage models with phoneme-to-word and feature-to-phoneme feedback connectivity which generated at least one contextual error. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	+	26.4	1243	< .001	*
joltPrimeRatio	−	3.5	12	< .001	*
decay	−	0.7	0	0.507	
steps	+	29.6	1772	< .001	*
actiNoiseSD	+	13.0	186	< .001	*
intrinNoiseSD	−	1.1	1	0.288	

Table 6.12: Results of logistic regression model analyses using parameter values to predict the occurrence of lexical bias and phonological similarity effects, for all one-stage models with both phoneme-to-word and feature-to-phoneme feedback connectivity which generated at least one contextual error. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	+	25.9	1679	< .001	*
joltPrimeRatio	−	2.8	8	0.004	*
decay	−	0.9	1	0.385	
steps	+	27.9	2089	< .001	*
actiNoiseSD	+	13.2	195	< .001	*
intrinNoiseSD	−	0.9	1	0.344	

these effects. This can be seen more clearly in graphs of the median error rates for the lexical and non-lexical outcome conditions (figure 6.12) and the median error rates for the phonologically similar and dissimilar conditions (figure 6.13). Whilst error rates are high at low jolt to prime ratios, and quite high at high activation-based noise levels, there is little difference between the number of lexical and non-lexical outcome errors generated. Similar statements can be made when comparing error rates in the phonologically similar and dissimilar conditions. This suggests that the main role of these parameters is to boost the power of the analysis. At high connectivity strength settings or where there is a high number of steps before selection however, whilst error rate increases overall, the median error rate is clearly higher for the lexical outcome condition and the phonologically similar condition. We argue that these parameters allow activation flow through feedback loops to have a greater influence on model behaviour.

Dell (1986) also argued on the basis of his simulations that lexical bias and phonological similarity effects become more pronounced when more steps pass before selection. We note that contrary to previous chapters, our findings regarding the effect of manipulating the steps parameter match Dell's (1986) in this regard. It is also worth highlighting that according to the feedback explanation of the lexical bias and phonological similarity effects, it is impossible for these effects to occur when only two timesteps are permitted before selection, as this is not enough time for activation to pass from the jolted word, to the phoneme, either upwards to the word level or downwards to the featural level, and then back to the phoneme; three timesteps would be the minimum required for activation to pass along these paths.

Finally, we consider which parameter settings allow the model to exhibit lexical bias effects and phonological similarity effects whilst observing the constraints on error rate and non-contextuality. High connectivity strengths, low jolt to prime ratios, high numbers of steps before selection, and high levels of activation-based noise all lead to higher error rates, and with the exception of low jolt to prime ratios, also cause higher proportions of non-contextual errors to be generated. Correspondingly, figure 6.14 suggests that medium values of these parameter settings lead to the largest numbers of specific models which exhibit both effects whilst passing both constraints.

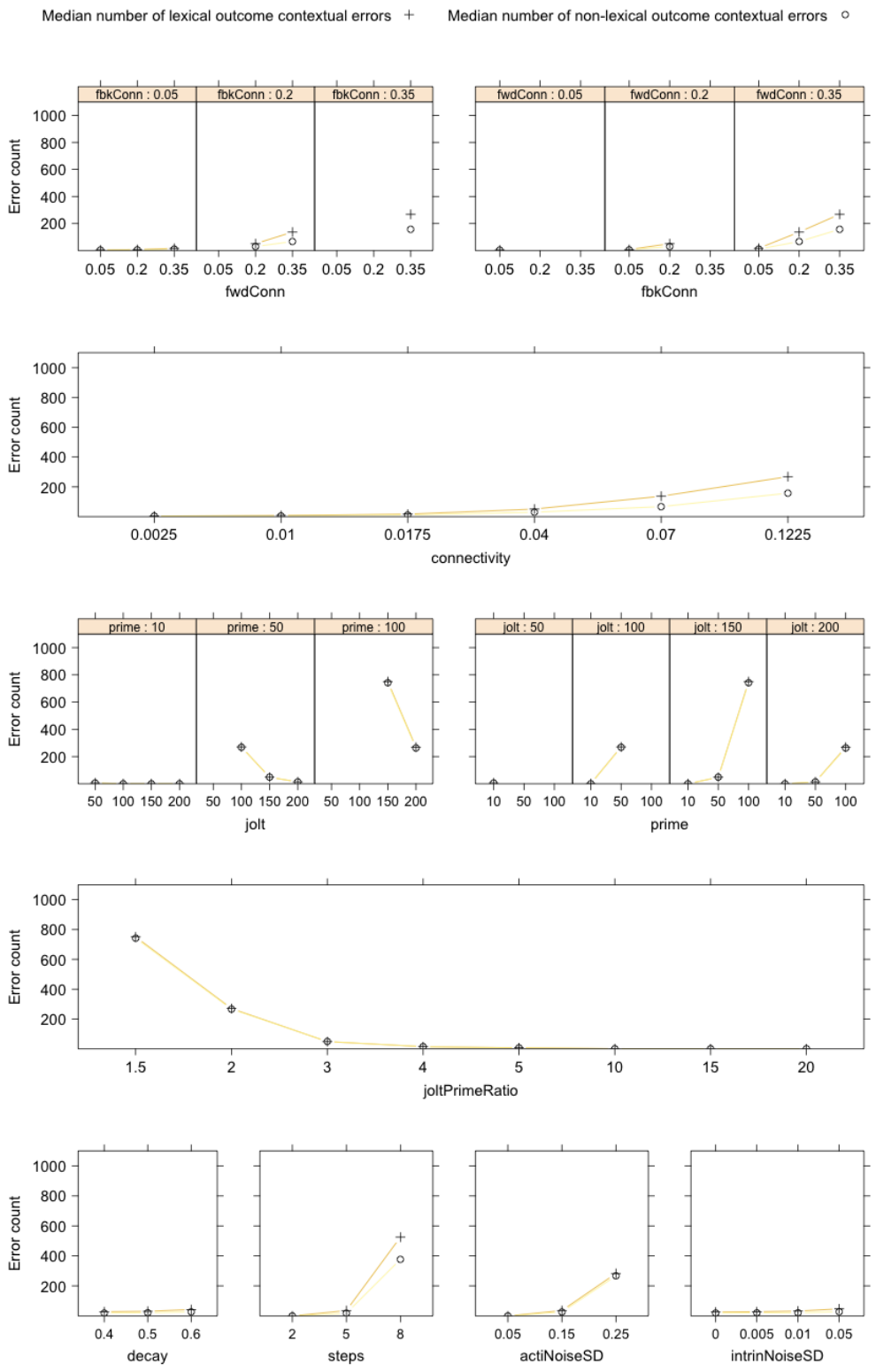


Figure 6.12: The effect of changing parameter settings on the median number of contextual errors generated in the lexical outcome condition, and the median number of contextual errors generated in the non-lexical outcome condition, for all one-stage phonological encoding models with phoneme-to-word and feature-to-phoneme feedback.

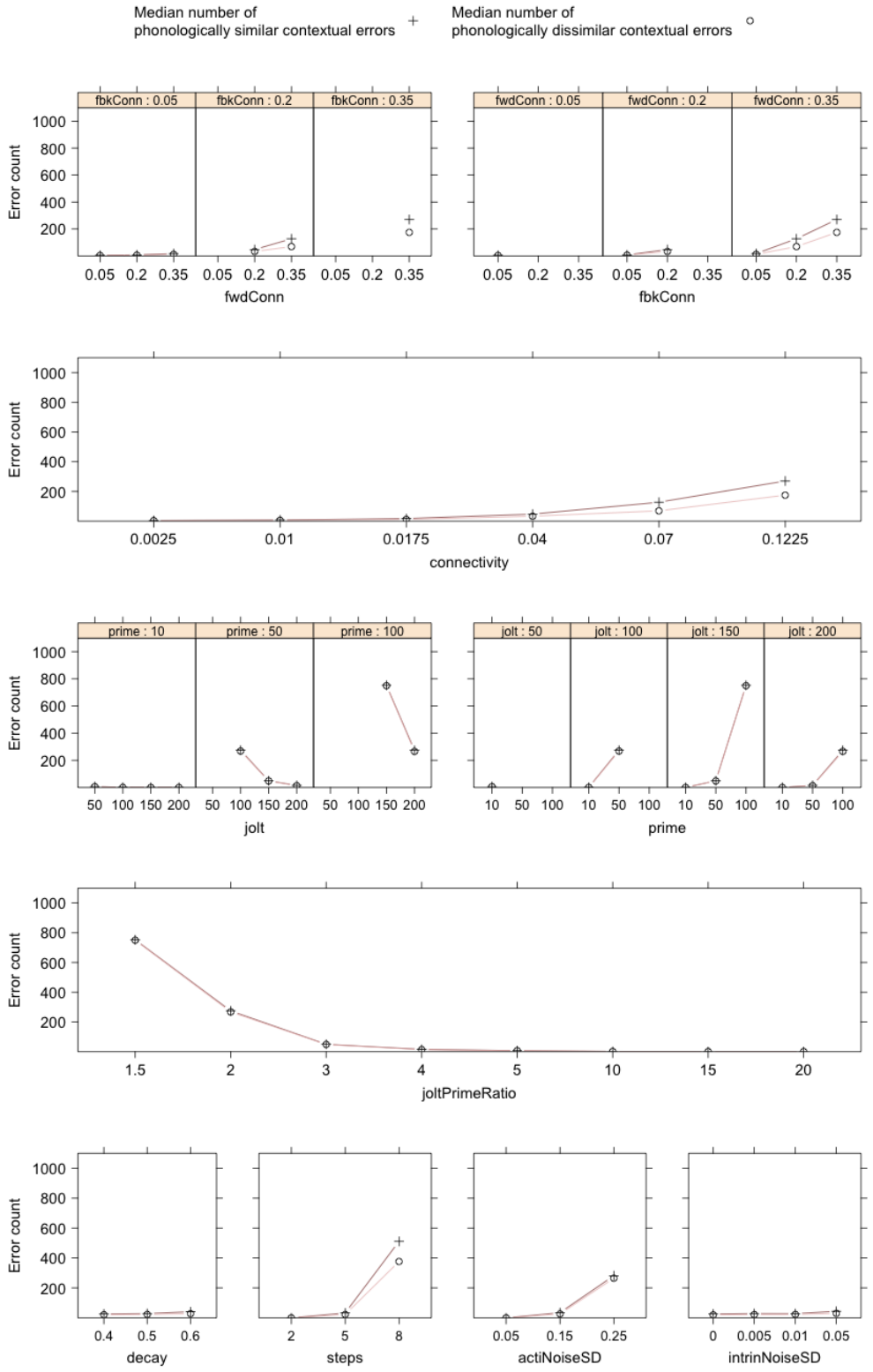


Figure 6.13: The effect of changing parameter settings on the median number of contextual errors generated in the phonologically similar outcome condition, and the median number of contextual errors generated in the phonologically dissimilar outcome condition, for all one-stage models with phoneme-to-word and feature-to-phoneme feedback.

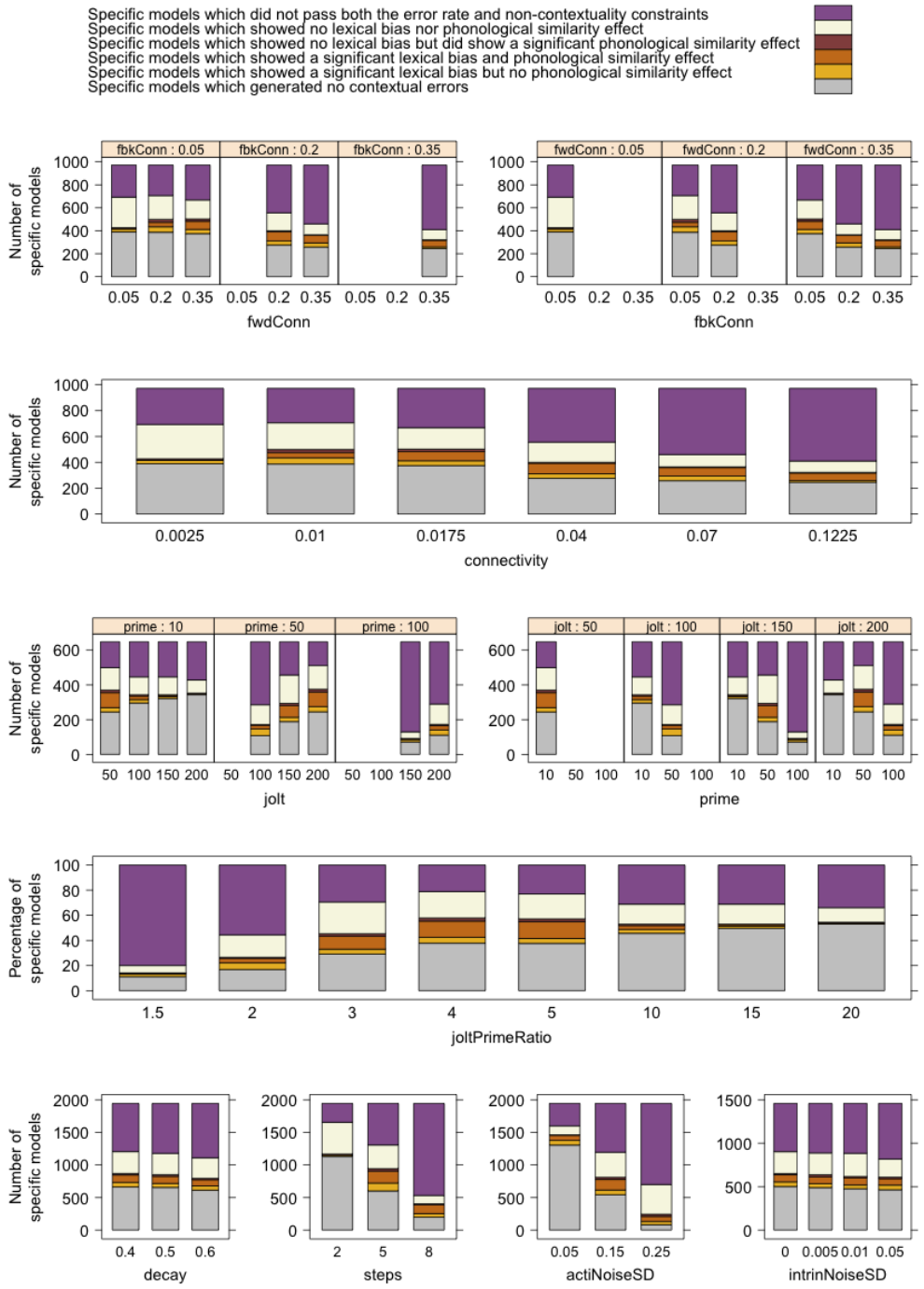


Figure 6.14: The effect of changing parameter settings on exhibition of lexical bias and phonological similarity effects in one-stage phonological encoding models with feature-to-phoneme feedback, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

6.5 Conclusions

In this chapter, we have introduced a binomial method for determining whether a model architecture allows a statistical pattern observed in human behaviour to be accounted for. Using the well established lexical bias effect and phonological similarity effect as test cases, we showed that in a model with output at the phoneme level, the binomial method indicates that feedback from phonemes to words is required to account for the lexical bias effect, and feedback from features to phonemes is required to account for the phonological similarity effect. We showed that these architectures are able to account for these effects whilst still observing the constraints on error rate and non-contextuality of errors.

We further demonstrated that the architecture with feedback from phonemes to words and from features to phonemes can account for both the lexical bias effect and phonological similarity without different parameter settings being required for the different effects, and also whilst observing the constraints on error rate and non-contextuality of errors. We noted however that the binomial method in its current form loses power as the number of effects which should be simultaneously accounted for increases. This is a methodological point which should be revisited by future research.

Finally, we used the methods developed in chapter 4 to determine which parameters are required for a lexical bias and phonological similarity effect to be displayed. We found that the same parameter settings are important for both effects, such that models with high connection strength, a low jolt to prime ratio, a high number of steps before selection and high levels of activation-based noise are most successful. We argued that some of these parameter settings support the lexical bias and phonological similarity effect by generally increasing error rate and therefore increasing the power of the logistic regression analysis carried out on each specific model, whereas some parameter settings specifically support activation flow through the feedback loops which underlie the lexical bias and phonological similarity effect.

In the following chapters, we will apply the methodology developed here to investigate the behaviour of model with output at the featural level. We will determine whether this model can account for old transcribed results and new instrumental evidence, and what these results tell us about activation flow between phonemes and features.

6.6 Chapter summary

In this chapter, we introduced a binomial method for determining whether a model architecture allows patterns observed in human behaviour to be accounted for, using the lexical bias effect and phonological similarity effect in a model with output at the phoneme level as test cases. We demonstrated how this method can be applied to determine whether an architecture can account for human results whilst observing the constraints on error rate and non-contextuality of errors determined in chapter 4. We further showed how methods developed in chapter 4 can elucidate the role of the parameter settings in an architecture's ability to account for a given finding. This method is applied throughout the rest of the thesis to determine what constraints old and new evidence places on activation flow between phonemes and features in a model with output at the featural level.

Chapter 7

Activation flow between phonemes and features: transcribed and acoustic measurements of categorised productions

7.1 Introduction

A key aim of this thesis was to clarify what current speech error evidence tells us about how activation flows between phonemes and subphonemic representations. In the second half of this thesis, we use the methodology and foundational understanding of the behaviour of Dell's (1986) model which we developed in the previous chapters to begin to address this problem directly. We consider classic transcribed speech error results, and newer instrumental findings.

In Dell's (1986) original model, it was assumed that all errors occur at the phoneme level or above. Output in this model was therefore at the phoneme level, such that subphonemic errors could not occur. However, in section 2.3.1, we argued that the evidence presented for an absence of subphonemic errors is not convincing in the light of findings from the perceptual literature. Instead, the results which led to this conclusion may have been unduly influenced by speech error collectors' perceptual systems. Furthermore, to model new instrumental evidence (e.g., Goldrick & Blumstein, 2006; McMillan, 2008; McMillan et al., 2009), the model must generate output below the phoneme level. Here, we therefore begin to examine the behaviour of a model with two processing stages: phonological encoding and subphonemic processing. In this model, output is measured from the subphonemic level, which in this implementation constitutes a layer of features.

In section 2.3.2, we laid out a number of predictions regarding the ability of different models of activation flow between phonemes and features to account for number of

Table 7.1: Predictions of the ability of different two-stage models of information flow to account for empirical data.

	Predictions					
	transcribed LB	transcribed PS	G&B 2006 traces	G&B 2006 trace LB	MMea 2009 delta LB	MM 2008 delta PS
Cascading from all Ws to Ps						
No cascading from Ps to Fs	×	✓	✓	×	×	×
Cascading from selected Ps to Fs	×	✓	✓	×	×	×
Cascading from all Ps to Fs	×	✓	✓	×	×	×
Feedback from Fs to Ps	×	✓	✓	×	×	✓
Feedback from Ps to Ws						
No cascading from Ps to Fs	✓	✓	✓	×	✓	×
Cascading from selected Ps to Fs	✓	✓	✓	✓	✓	×
Cascading from all Ps to Fs	✓	✓	✓	✓	✓	×
Feedback from Fs to Ps	✓	✓	✓	✓	✓	✓

Key:
LB = lexical bias, PS = phonological similarity, G&B 2006 = Goldrick and Blumstein (2006), MM 2008 = McMillan (2008), MMea 2009 = McMillan et al. (2009)

Ws = words, Ps = phonemes, Fs = features

✓ = predicted to be able to account for evidence
× = predicted not to be able to account for evidence

Grey boxes indicate that our prediction does not match the standard claim in the literature.

speech error results. We present these predictions again in table 7.1. This table is reformatted to explicitly include predictions of the model's behaviour with and without feedback from phonemes to words. As this table illustrates, eight two-stage models will be tested, where the presence of phoneme-to-word feedback is varied orthogonally with the four phoneme-to-feature connectivity options.

In this chapter, we begin to report simulations seeking to verify these predictions. The studies reported here first focus on old evidence acquired via the transcription of speech errors: the lexical bias and phonological similarity effects. We then consider new evidence in which acoustic properties of productions classified as correct or erroneous are compared (Goldrick & Blumstein, 2006). In the final simulation chapter, we will consider evidence which does not rely on error categorisation by transcribers at all (McMillan, 2008; McMillan et al., 2009).

We begin by considering the transcribed lexical bias and phonological similarity effect. In section 2.3.2, we predicted that evaluating output at the feature level would have no effect on the model's ability to account for the lexical bias effect,

such that all models with feedback from phonemes to words would be able to explain this evidence. However, we argued that feedback from features to phonemes would not be required to account for the phonological similarity effect when output is at the featural level. As errors involving misselection of one feature should be more common than errors involving misselection of two features, any architecture should be able to explain this result.

We then examine the constraints imposed on the model by the findings reported by Goldrick and Blumstein (2006). Goldrick and Blumstein (2006) compared the VOTs of correct and erroneous onset consonant productions in tongue twisters, and found that an influence of the intended phoneme could be found on erroneous productions. For example, where an intended /k/ is produced as a [g], the resulting /g/ would be more voiceless than an intended and correctly produced [g]. Goldrick and Blumstein (2006) claimed that this result provided evidence that activation from the intended but unselected /k/ must cascade to the subphonemic level, and that cascading from all phonemes would be required to explain this result.

However, in section 2.3.2, we argued a model with no cascading from phonemes to features could also account for this result. It was suggested that a production transcribed as an incorrectly selected phoneme may in fact reflect a correctly selected phoneme which has been affected by noise at the featural level. For example, the intended /k/ may have been correctly selected, but during subphonemic processing, the voiced feature may have become more activated than the voiceless feature, resulting in a [g] production. In this case however, it is likely that the voiced feature would be less activated and the voiceless feature more activated than in a situation where a /g/ was selected at the phoneme level. A trace of the intended /k/ would therefore be evident on the unintentional [g] production, with no cascading from phonemes required.

We also claimed that a further mechanism for trace generation may operate in architectures with cascading from selected phonemes. We argued that intentionally selected phonemes will be more strongly activated than unintentionally selected phonemes, and that as a result, voicing characteristics at the featural level will be more weakly activated where the phoneme was not intentionally selected. For example, where a /k/ was intended but a /g/ selected, the unintended and more weakly activated /g/ will pass less activation to the voiced feature. By definition, a less voiced production is more voiceless, such that a trace of the intended /k/ would

be present in the final articulation, with no cascading from unselected phonemes required.[1]

Finally, we consider what model architecture is required to account for Goldrick and Blumstein's (2006) post-hoc result that traces of intended phonemes on unintended productions of competing phonemes were smaller when error outcomes were lexical rather than non-lexical. For example, a smaller trace of the intended /k/ would be found on a production of /g/ in the error, *"kess"* → *"guess"* than in the error *"keff"* → *"geff"*. Goldrick and Blumstein (2006) claimed that this result reflected suppression of the activation cascading from the intended phoneme due to greater activation of the unintended phoneme in the lexical error outcome condition. This explanation would require that activation cascades from unselected phonemes. However, we argued that extra activation cascading from the unintentionally selected phoneme in the lexical error outcome condition is sufficient to explain this result, as stronger activation of the unintended voicing characteristics would diminish the trace of the intended phoneme. According to this theory, any architecture with cascading from selected phonemes would be able to account for Goldrick and Blumstein's (2006) post-hoc finding.

In the next sections, we first consider the basic behaviour of different two-stage model architectures, reporting on error rate and non-contextuality of errors. We then present results from simulations designed to evaluate the predictions laid out above.

7.2 Error rate and non-contextuality of errors in two-stage models

In this section, we look at the basic behaviour of two-stage models.

7.2.1 Simulation methodology

To determine how architecture manipulations affect the basic behaviour of two-stage models and discover which specific models display error rates and proportions of non-contextual errors in line with the limits we prescribed in chapter 4, we ran random word production simulations like those reported in chapter 4, with

[1] We do not believe that the use of two features to represent voicing is necessary for this argument to be valid. The main crux of the argument is that VOT is a one-dimensional measure, such that productions which are less voiced are more voiceless, and vice-versa. We therefore believe that if only one voicing node was used, and voiced consonants activated this node whereas voiceless consonants inhibited it (or vice-versa), the same prediction would hold.

Table 7.2: Activation flow characteristics of the four proposed models of information flow between phonological encoding and subphonemic processes (replicated from table 2.2)

Model	**Information from phonological encoding**			**Feedback from subphonemic representations**
	Identity of selected phoneme	*Activation from selected phoneme*	*Activation from unselected phonemes*	
No casc	✓			
Casc from sel	✓	✓		
Casc from all	✓	✓	✓	
Feedback	✓	✓	✓	✓

adaptations for a two-stage model. The details of these simulations are outlined here.

Model configuration

We considered the behaviour of eight two-stage models of phonological encoding and subphonemic processing, by varying the presence of phoneme-to-word feedback orthogonally with four options for phoneme to feature connectivity: *no cascading from phonemes*, *cascading from selected phonemes only*, *cascading from all phonemes*, and *feedback from subphonemic representations*. Properties of these different connectivity options are recapped in table 7.2, and further details of their implementation are provided in chapter 3.

Parameter settings were varied for each architecture as outlined in section 3.6. Again, in architectures which contain no feedback from phonemes to words and no feedback from features to phonemes, the strength of feedback connectivity *fbkConn* is not varied. Combining architectures and parameter settings, a total of 37,908 specific two-stage models were tested.

Model task and lexicon

The same lexicon was used as for the simulations described in chapter 4, as well as exactly the same list of random word pairs for production. Again, we focused only on first word productions, but the second word in the word pair was primed.

Model output interpretation

Output from the two-stage models was classified as the phoneme formed from the most activated features in each syllable position at the end of subphonemic processing. Feature combinations which did not form a phoneme in the English phoneme

inventory were recorded but were not assigned a phonemic category. Onset productions were then again further classified as *correct productions*, if the intended onset was produced; *contextual errors*, if the competing onset was produced; and *non-contextual errors* if any other onset was produced, or if a feature combination not assigned a phonemic category was produced.

7.2.2 Simulation results

Figure 7.1 shows that the error rate for two-stage models tended to be higher than it was for one-stage models (compare to figure 6.1). Whereas the median error rate for all one-stage models was 0.29%, the median error rate for two-stage models was 0.56%. Figure 7.2 similarly shows that the proportion of non-contextual errors generated was higher for two-stage models than it was for one-stage models (compare to figure 6.2), with a median proportion of 0.65% non-contextual errors, compared to 0% for one-stage models. We suggest that these results are due to the fact that there are now two possible locations for errors to occur. Furthermore, misselection of one feature is likely to result in a non-contextual error, even if the other two features are correctly selected.

Clear increases in both error rate and non-contextuality of errors resulted from addition of feedback at either level, but particularly between features and phonemes. This is due to feedback connections allowing activation to flow to representations which do not form part of the intended utterance. The effect of feature-to-phoneme feedback is particularly strong on the proportion of errors which are non-contextual. This follows from the fact that feature-to-phoneme feedback will convey more activation to unrelated features. As noted above, misselection of a single feature will frequently lead to production of a non-contextual error.

Finally, both graphs also show that the error rate and proportion of non-contextual errors generated by models with feedback from phonemes to words, and cascading from all phonemes, was notably higher than it was for other architectures without feedback from features to phonemes, including the architecture with no feedback from phonemes to words and cascading from all phonemes. We suggest that in this architecture, feedback from phonemes to words both activates representations which do not form part of the intended utterance, and allows the prime activation to be reinforced where it would otherwise decay. Activation from primed and unrelated phonemes is then able to cascade to the featural level, causing more errors to be generated, particularly non-contextual errors.

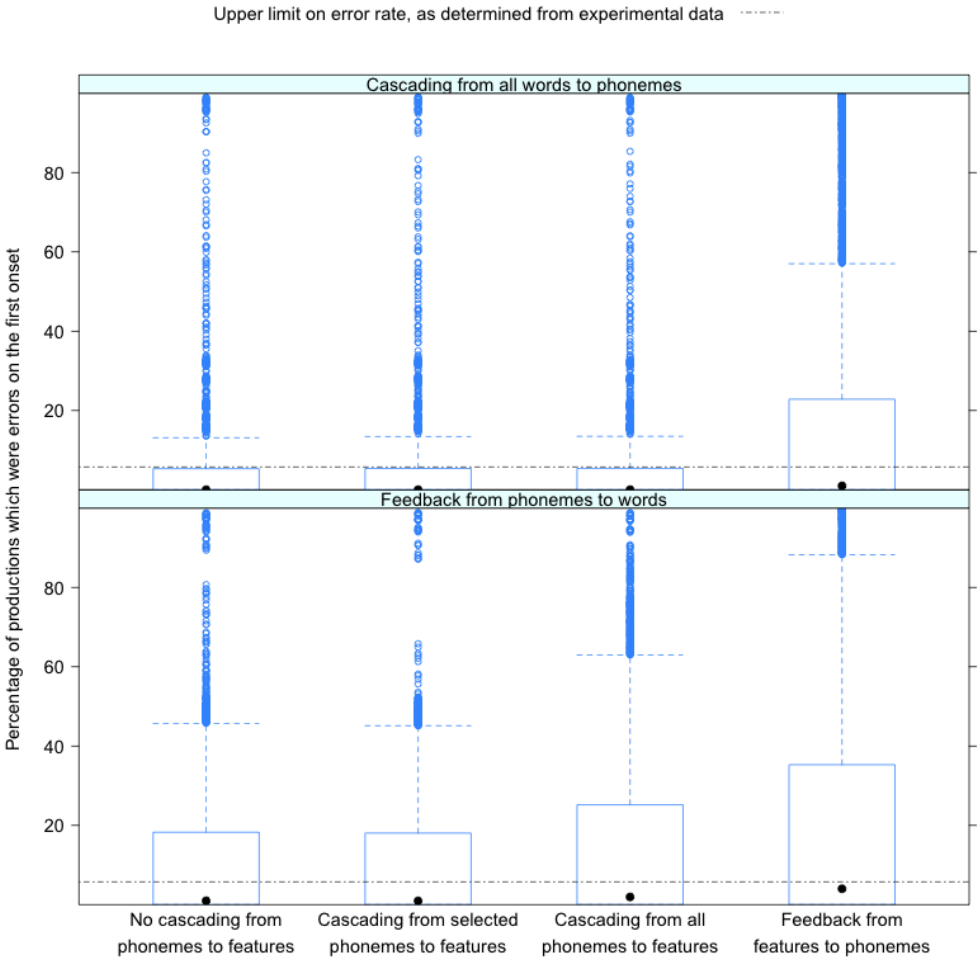


Figure 7.1: The effect of modifying word-to-phoneme and phoneme-to-feature activation flow on first onset error rate in two-stage models of phonological encoding and subphonemic processing. The dotted line represents the upper limit on error rate as calculated in chapter 4.

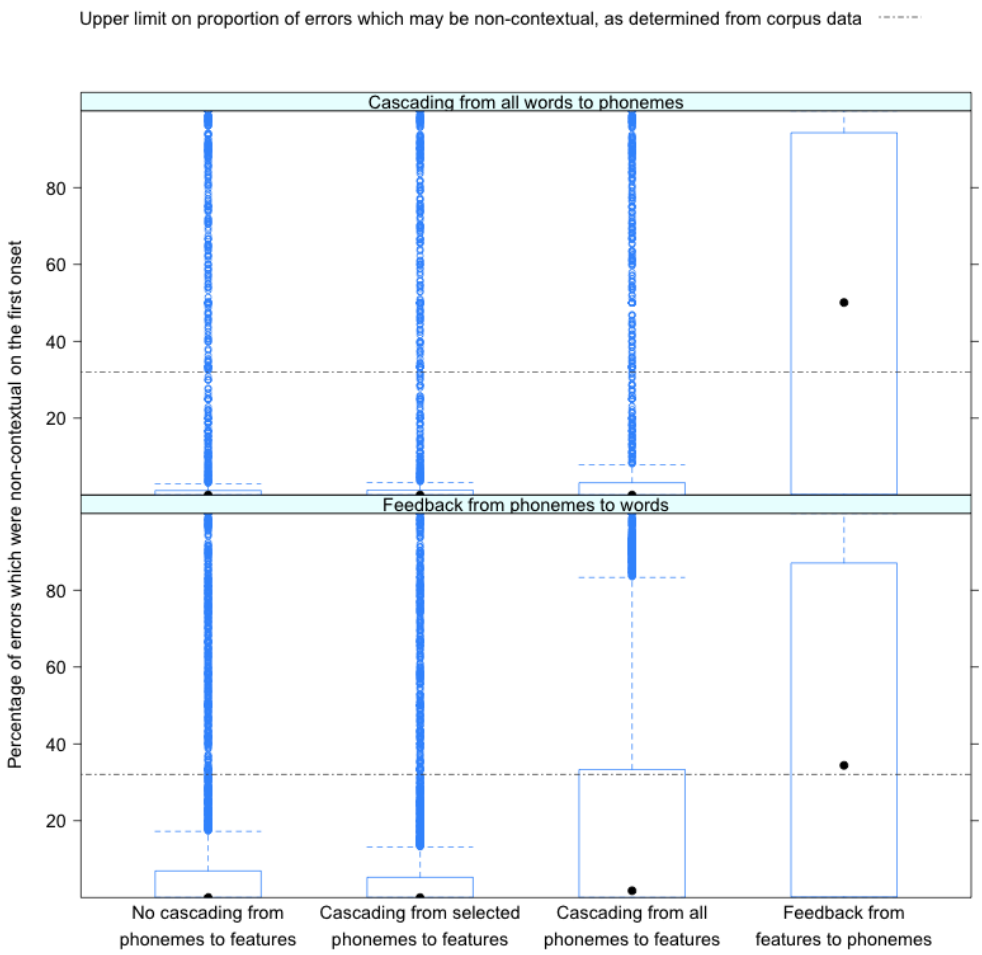


Figure 7.2: The effect of modifying word-to-phoneme and phoneme-to-feature activation flow on the proportion of errors which are non-contextual at the first onset in two-stage models of phonological encoding and subphonemic processing. This proportion can only be calculated for specific models which generated at least one error. The dotted line represents the upper limit on error non-contextuality as calculated in chapter 4.

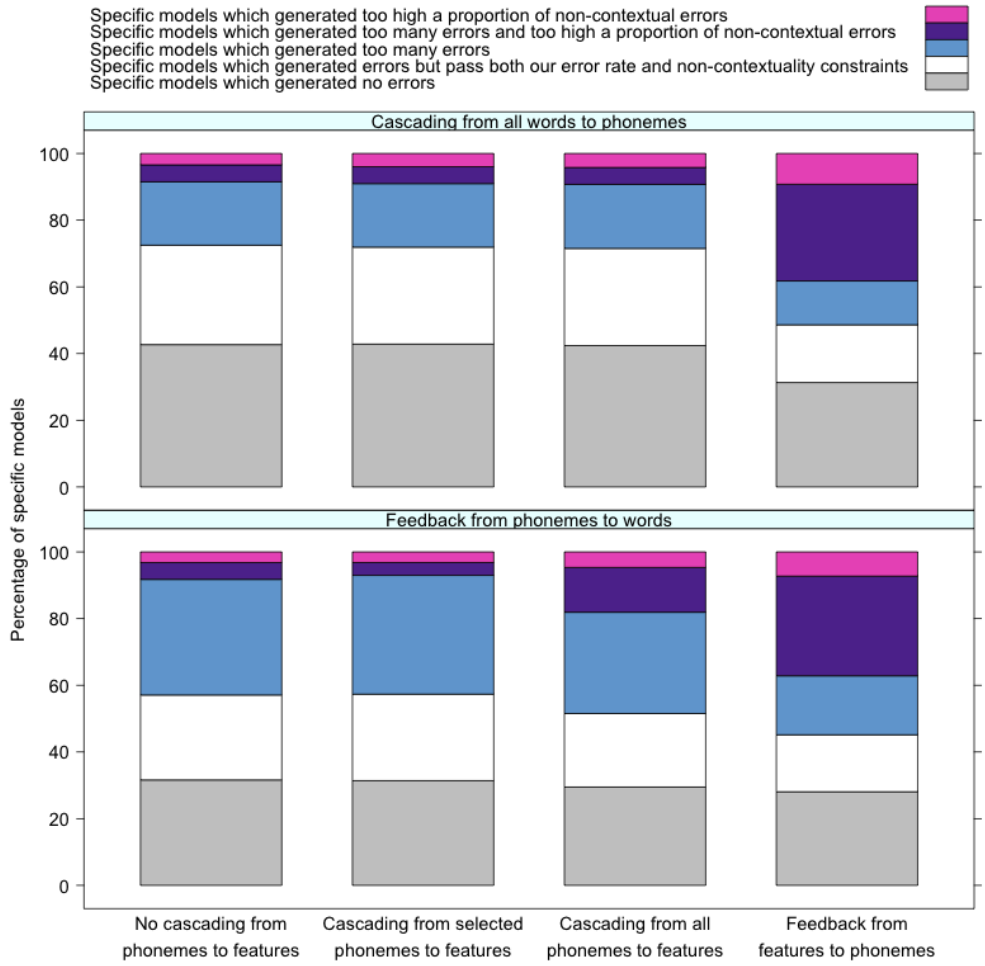


Figure 7.3: The effect of modifying word-to-phoneme and phoneme-to-feature activation flow on the numbers of specific models which pass our constraints on error rate and error non-contextuality, for all two-stage models. See figure 4.9 in chapter 4 for further elaboration on the key.

Figure 7.3 shows that the number of specific models which passed the constraints for each architecture reflects the observations made above. As argued in chapter 2, transcribers may not hear all featural errors, many of which may result in non-contextual errors. In the current simulation, featural errors cannot be missed as the output of the current models is determined by directly reading the activation levels of the feature nodes. However, both the upper limit on error rate and the upper limit on non-contextuality were designed to be overestimates. We therefore continue to use these limits as a guide to what constitutes reasonable behaviour of the network, until a better source of empirical data becomes available.

To investigate possible differences in the effects of parameter manipulations on model behaviour between one-stage and two-stage models, we investigated the behaviour of the two-stage architecture with feedback from phonemes to words and feedback from features to phonemes in more detail. Table 7.3, figure 7.4 and figure 7.5 show that in this architecture, the effects of manipulating spreading activation parameters on first word production error rate and non-contextuality are similar to the effects we reported for Dell's (1986) original one-stage model in chapter 4. A key difference however is that in this two-stage model, more errors and a higher proportion of non-contextual errors are now generated with low rather than high decay rates. We argue that in the two-stage model, a higher decay rate results in a cleaner network following the first stage, such that the activation signal from phonological encoding is clearer. Our results suggest that at low decay levels, activation on unrelated representations accumulates during the first processing stage and disturbs the second processing stage.

Figure 7.5 illuminates a strange effect, such that the number of non-contextual errors is very high when intrinsic noise is absent from the network, although once intrinsic noise is present, higher levels of this noise appear to cause higher proportions of non-contextual productions. This is reflected in the logistic regression summarised in table 7.3, where the effect of intrinsic noise on the proportion of non-contextual errors has a negative direction. It is currently unclear what brings this behaviour about.

These results are all reflected in figure 7.6, which shows how many specific models at each parameter setting pass the constraints on error rates and non-contextuality of errors.

7.2.3 Conclusions

In this section, we showed that two-stage models generally demonstrate higher error rates and proportions of non-contextual errors than one-stage models. We found that two-stage models with feedback at any point in the model demonstrate higher error rates and proportions of non-contextual errors than two-stage models without feedback. A similar result was found for one-stage models. However, the architecture with feedback from phonemes to words and cascading from all phonemes also exhibits higher error rates and proportions of non-contextual errors than other architectures with no feedback from phonemes to features. Finally, we demonstrated that for the architecture with feedback from phonemes to words and from features to

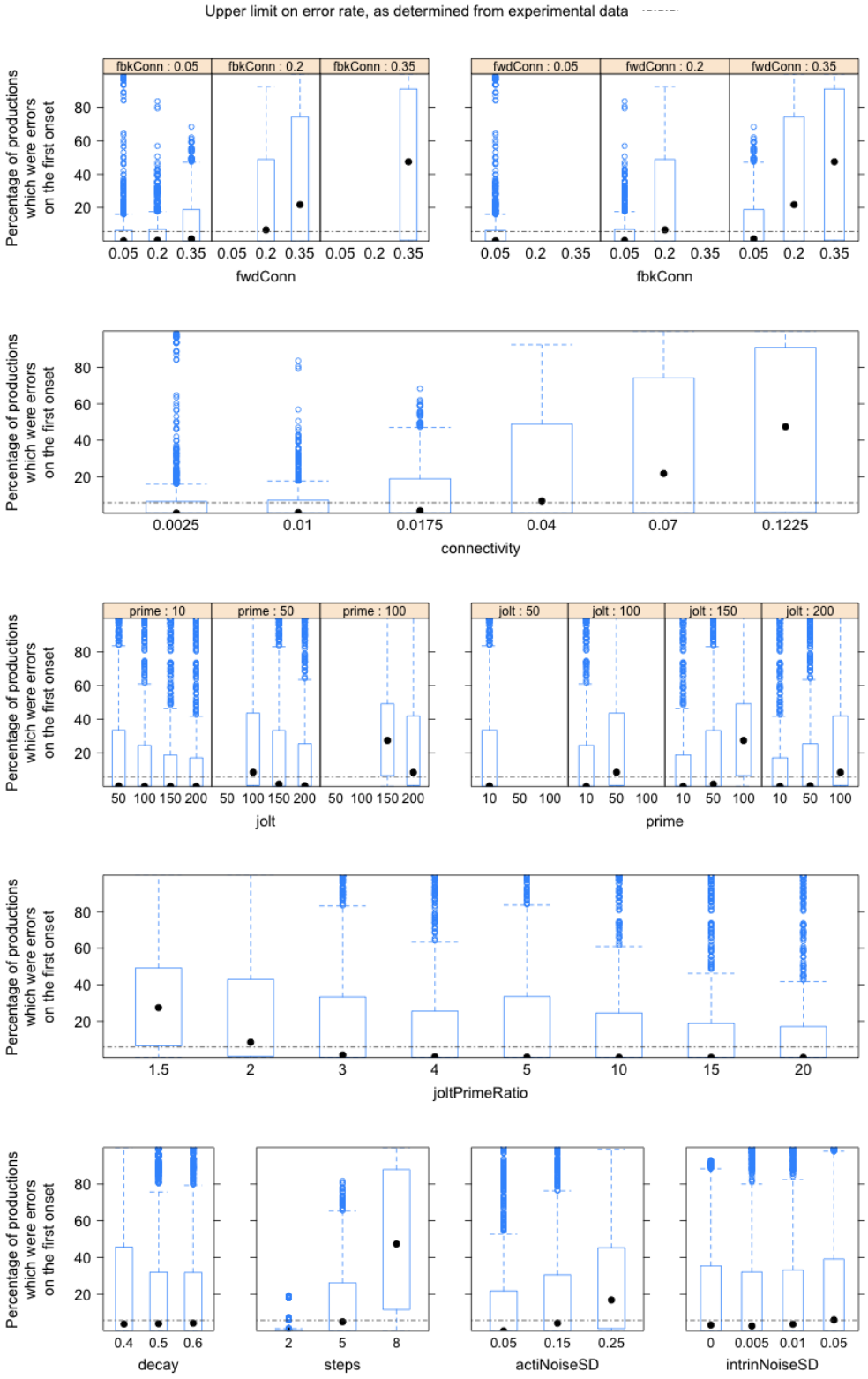


Figure 7.4: The effect of changing parameter settings on first onset error rate, for all specific two-stage models with phoneme-to-word and feature-to-phoneme feedback. The dotted line represents the upper limit on error rate as calculated in chapter 4.

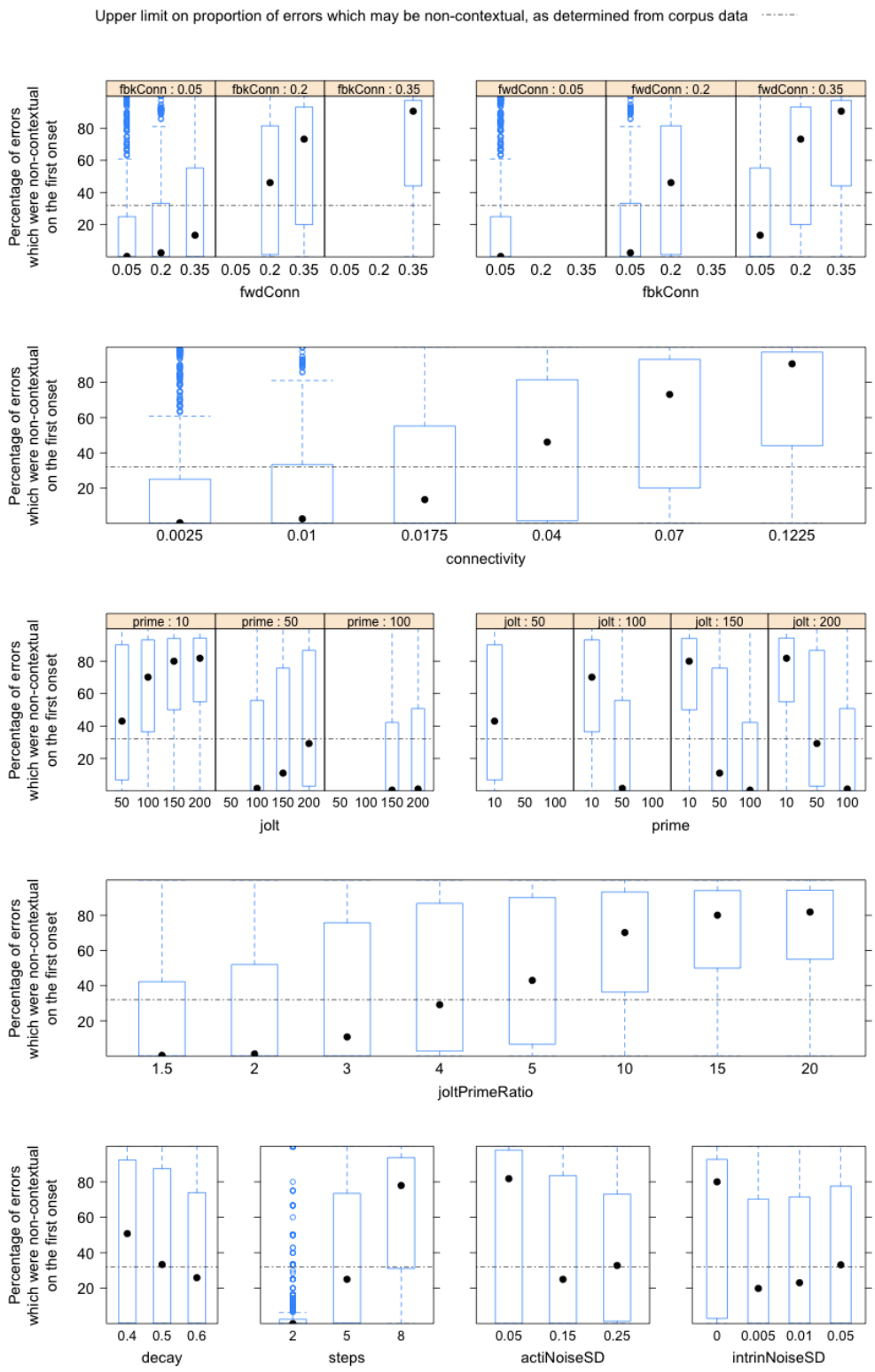


Figure 7.5: The effect of changing parameter settings on the proportion of errors which are non-contextual at the first onset, for specific two-stage models with phoneme-to-word and feature-to-phoneme feedback. This proportion can only be calculated for specific models which generated at least one error. The dotted line represents the upper limit on error non-contextuality as calculated in chapter 4.

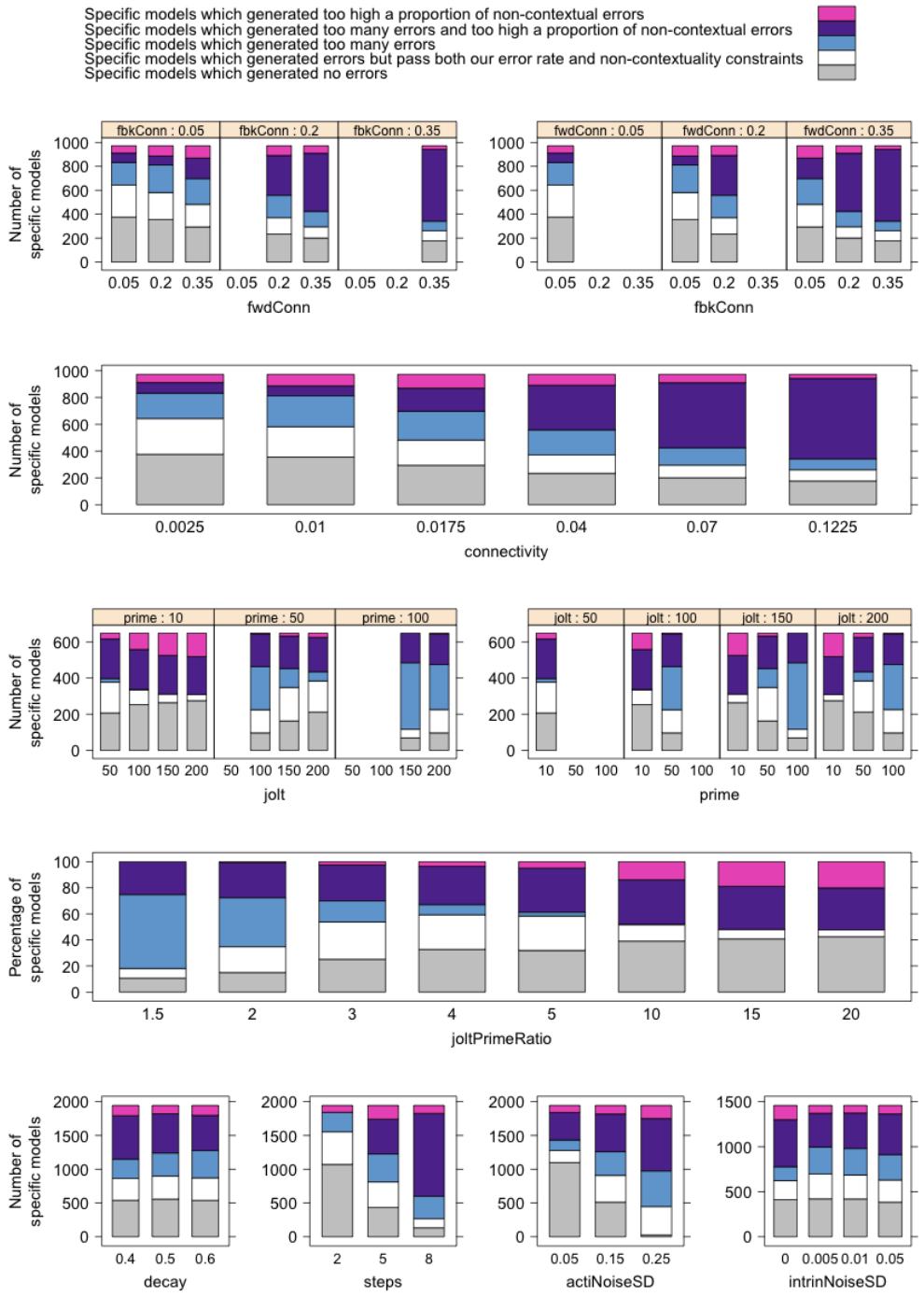


Figure 7.6: The effect of changing parameter settings on the numbers of specific models which pass our constraints, for specific two-stage models with feedback from phonemes to words and from features to phonemes. See figure 4.9 in chapter 4 for further elaboration on the key.

Table 7.3: Results of logistic regression model analyses using parameter values to predict error rate and proportion of errors which were non-contextual on the first onset for two-stage models with phoneme-to-word and feature-to-phoneme feedback. The proportion of errors which were non-contextual can only be calculated for specific models which generated at least one error on the specified onset. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	**Error rate**					**Non-contextuality**				
	Dir	Z	LRT	P (χ^2)		Dir	Z	LRT	P (χ^2)	
connectivity	+	2773.4	11328245	$< .001$	*	+	1273.5	2206948	$< .001$	*
joltPrimeRatio	−	871.1	817048	$< .001$	*	+	565.2	384589	$< .001$	*
decay	−	542.9	298621	$< .001$	*	−	620.9	402600	$< .001$	*
steps	+	3013.4	18604370	$< .001$	*	+	1181.9	1861555	$< .001$	*
actiNoiseSD	+	970.5	981616	$< .001$	*	+	97.4	9501	$< .001$	*
intrinNoiseSD	+	201.9	40549	$< .001$	*	−	110.0	12037	$< .001$	*

Key:
Dir = direction

phonemes, the effects of manipulating the spreading activation parameters on first word production error rate and non-contextuality are very similar to those that we reported for productions on the first word in Dell's (1986) original one-stage model in chapter 4, with the exception that in this two-stage architecture, more errors and higher proportions of non-contextual errors are generated at low decay rates.

7.3 The classic lexical bias and phonological similarity effects

In this section, we give details of the simulations used to investigate firstly which two-stage architectures can account for the lexical bias and phonological similarity effects as reported in speech error investigations that relied on transcription, and secondly which spreading activation parameter settings are required for these effects to be found. We then report the results of our simulations.

7.3.1 Simulation methodology

To examine the ability of various two-stage architectures to account for the lexical bias and phonological similarity effects, we carried out simulations very similar to those reported in chapter 6, again with adaptations for a two-stage model. The details of these simulations are outlined here.

Model configuration

We examined the behaviour of all two-stage models, leading us to test 37,908 specific models in total.

Model task and lexicon

We used the same lexicon and materials as described in chapter 6, considering single word production while a competitor word was primed. Again, we ran one simulation using the materials in which the place of articulation of the competitor always differs, and another in which the voicing feature of the competitor always differs. We report results from the simulation in which the place of articulation of the competitor always differs, noting any points where results from the other simulation were not the same.

Model output interpretation

Output from the two-stage models was classified as for the error rate and non-contextuality simulations described in section 7.2. Logistic regressions of the effects of lexicality and phonological similarity on contextual error generation were then carried out in the same way as in chapter 6.

7.3.2 Simulation results

We first investigate which architectures are able to generate the lexical bias effect, and which display a phonological similarity effect, as well as which architectures are able to account for both results. We then look at the effect of parameter manipulations on whether models display these effects.

Activation flow options required for lexical bias and phonological similarity

In this section, we investigate how activation must flow between words, phonemes and features for the lexical bias and phonological similarity effects to be displayed by two-stage models.

As for one-stage models, figure 7.7 and the binomial analysis summarised in table 7.4 show that plenty of evidence was found for the ability of models with feedback from phonemes to words to generate the lexical bias effect, but that no evidence was found for this ability in other models. Again, this conclusion still holds when specific

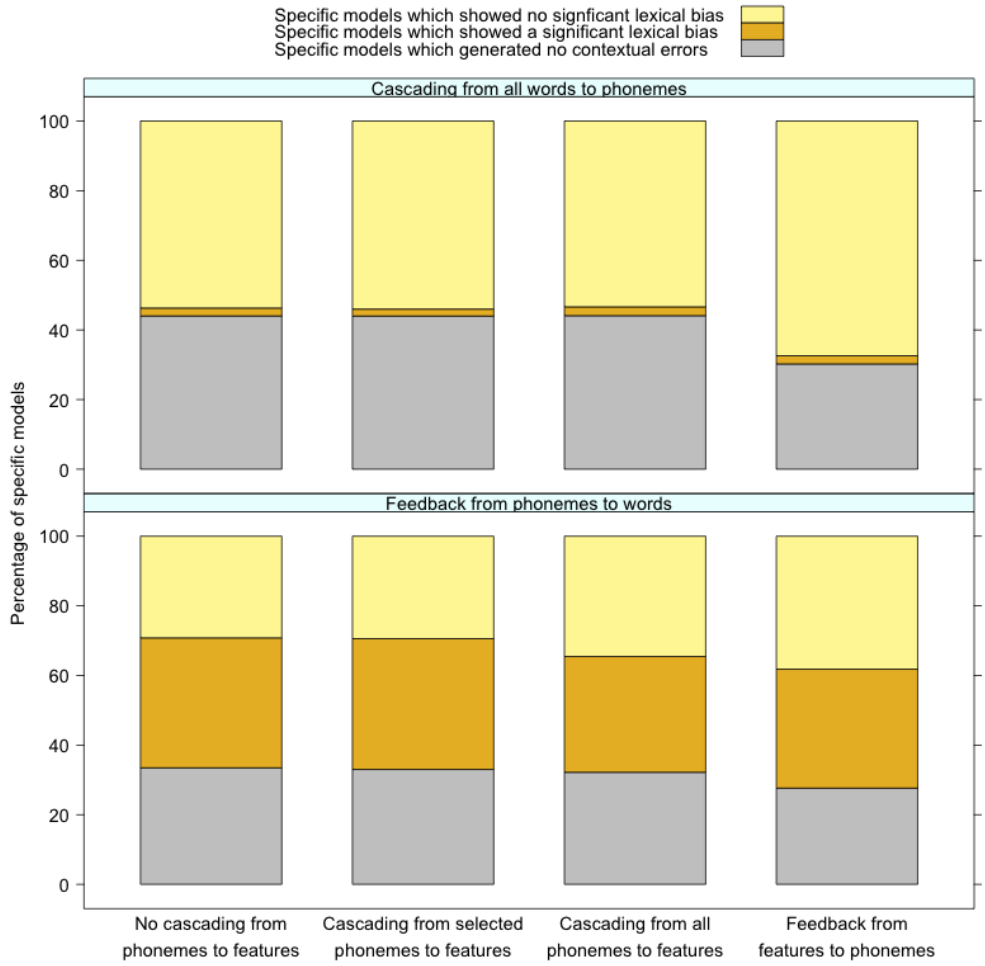


Figure 7.7: The effect of modifying word-to-phoneme and phoneme-to-feature activation flow on exhibition of lexical bias effects in two-stage models of phonological encoding and subphonemic processing.

models which fail the error rate and non-contextuality constraints are excluded, as shown in figure 7.8 and table 7.5.

As predicted however, feedback from features to phonemes was not required for two-stage models to exhibit a phonological similarity effect. Figure 7.9 and table 7.6 show that all of the tested architectures exhibit this effect. Again, we argue that the effect can be explained simply by reference to the fact that less noise is required to misselect one feature than two, without any claims about phoneme to feature connectivity. (This is the one result for which the results from the material set in which voicing always differed between target and competitor voicing results were slightly different. In this result set, the number of models displaying significant phonological simialrity results found for the architectures with feedback

Table 7.4: Binomial analysis to determine which two-stage architectures can generate a lexical bias effect. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating significant lexical bias effects by chance.

	Specific model counts			**Prob.**	
	Total	Generated contextual errors	Significant lexical bias		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	1632	66	$> .9$	
Cascading from selected Ps to Fs	2916	1633	58	$> .9$	
Cascading from all Ps to Fs	2916	1630	74	0.785	
Feedback from Fs to Ps	5832	4069	137	$> .9$	
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	3880	2177	$< .001$	*
Cascading from selected Ps to Fs	5832	3908	2191	$< .001$	*
Cascading from all Ps to Fs	5832	3955	1939	$< .001$	*
Feedback from Fs to Ps	5832	4220	1992	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

Table 7.5: Binomial analysis to determine which two-stage architectures can generate a lexical bias effect, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating significant lexical bias effects by chance.

	Specific model counts				**Prob.**	
	Total	Excluded	Generated contextual errors	Significant lexical bias		
Cascading from all Ws to Ps						
No cascading from Ps to Fs	2916	784	848	31	$> .9$	
Cascading from selected Ps to Fs	2916	801	832	26	$> .9$	
Cascading from all Ps to Fs	2916	808	822	27	$> .9$	
Feedback from Fs to Ps	5832	2963	1106	40	$> .9$	
Feedback from Ps to Ws						
No cascading from Ps to Fs	5832	2470	1410	439	$< .001$	*
Cascading from selected Ps to Fs	5832	2467	1441	442	$< .001$	*
Cascading from all Ps to Fs	5832	2756	1199	239	$< .001$	*
Feedback from Fs to Ps	5832	3165	1055	211	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

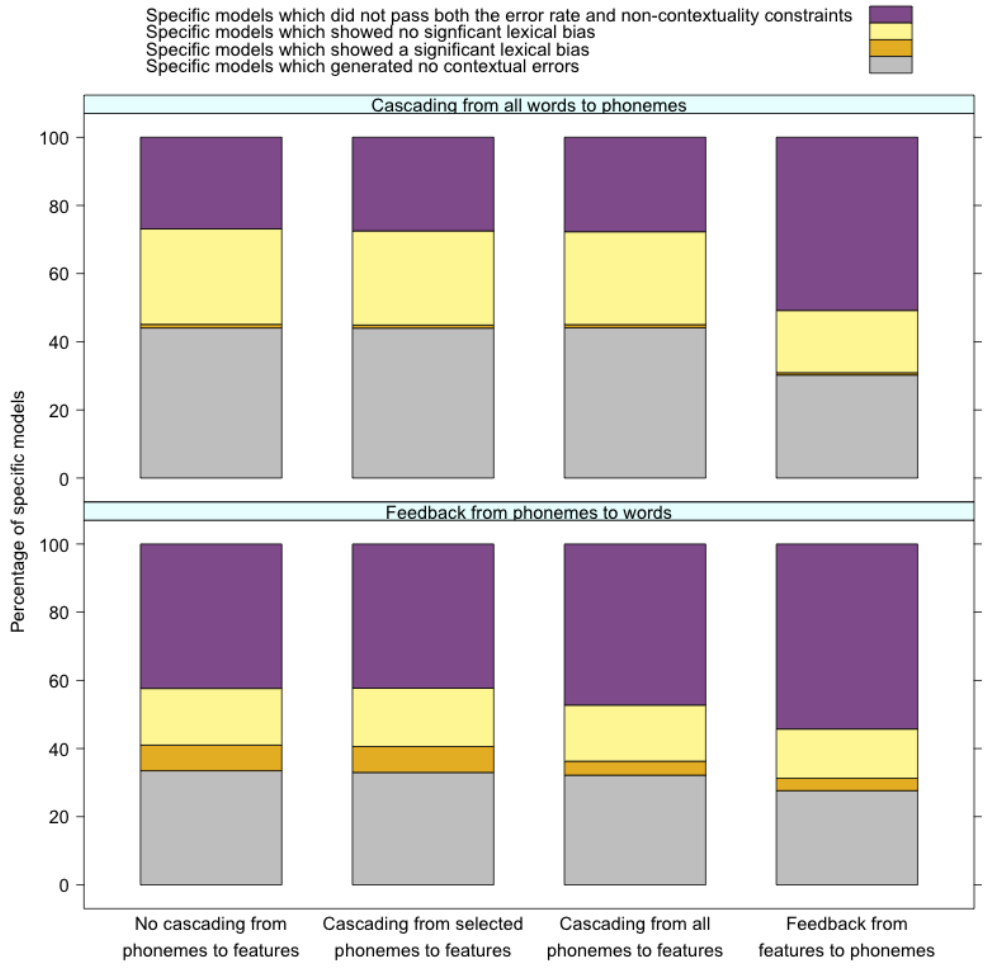


Figure 7.8: The effect of modifying word-to-phoneme and phoneme-to-feature activation flow on exhibition of lexical bias effects in two-stage models of phonological encoding and subphonemic processing, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

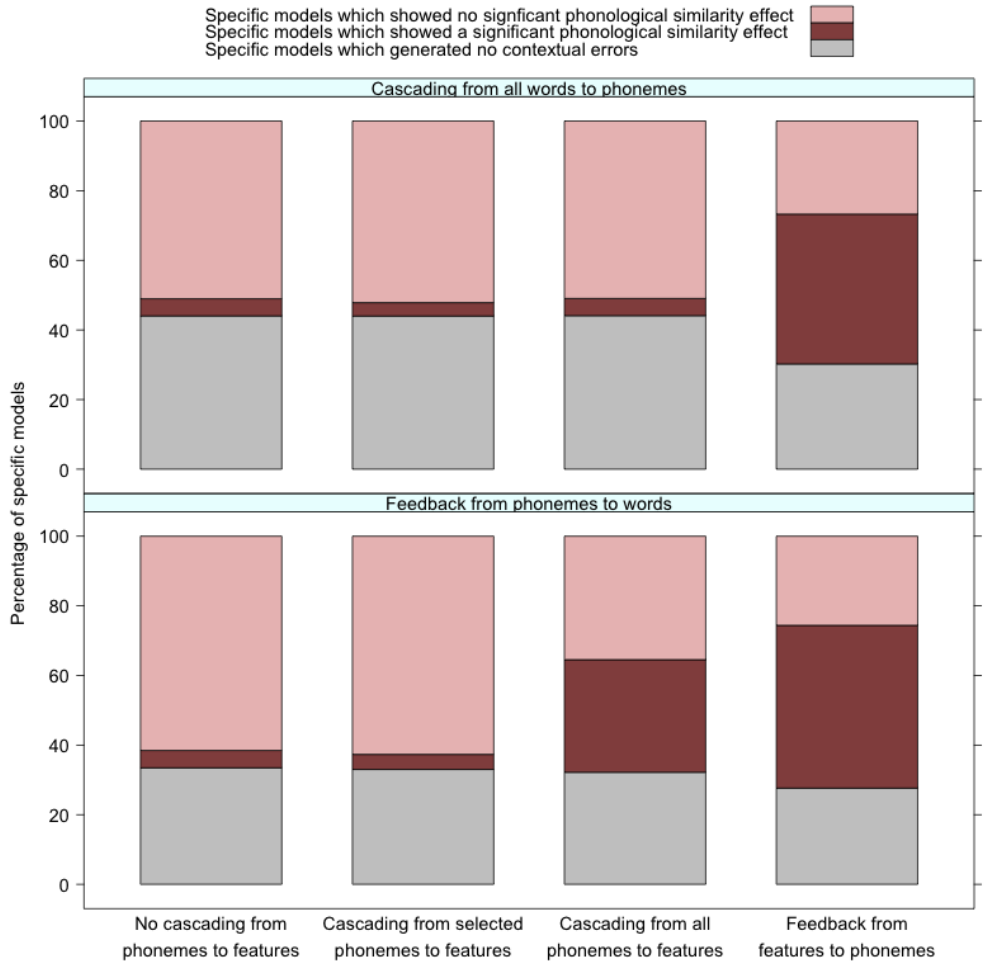


Figure 7.9: The effect of modifying word-to-phoneme and phoneme-to-feature activation flow on exhibition of phonological similarity effects in two-stage models of phonological encoding and subphonemic processing.

from phonemes to words, and either no cascading from phonemes to features, or cascading from selected representations, was not greater than would be predicted by chance. All other results were however the same.) However, it is clear from figure 7.9 that many more models exhibit phonological similarity effects for the architecture with feedback from phonemes to words and cascading from all phonemes to features, and the two architectures with feedback from features to phonemes, than for any of the remaining five architectures.

Excluding specific models which failed either or both of the constraints on error rate or non-contextuality causes problems for these same five architectures. Whilst figure 7.10 and table 7.7 show that there is still clear evidence that the phonological similarity effect is accounted for by the architecture with feedback from phonemes

Table 7.6: Binomial analysis to determine which two-stage architectures can generate a phonological similarity effect. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating significant phonological similarity effects by chance.

	Specific model counts			**Prob.**	
	Total	Generated contextual errors	Significant phonological similarity effect		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	1632	143	$< .001$	*
Cascading from selected Ps to Fs	2916	1633	112	$< .001$	*
Cascading from all Ps to Fs	2916	1630	144	$< .001$	*
Feedback from Fs to Ps	5832	4069	2510	$< .001$	*
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	3880	290	$< .001$	*
Cascading from selected Ps to Fs	5832	3908	253	$< .001$	*
Cascading from all Ps to Fs	5832	3955	1887	$< .001$	*
Feedback from Fs to Ps	5832	4220	2727	$< .001$	*

Key:
Ws = words, Ps = phonemes, Fs = features
Prob. = probability

to words and cascading from all phonemes to features, and the architectures with feedback from features to phonemes, no evidence of the phonological similarity effect is found for other architectures. (Results from simulations using the materials in which voicing always differs between target and competitor show exactly the same pattern.)

We argue that the problems experienced by these five architectures are due to difficulty in contextual error generation, a problem which occurs due to an implementation decision in our simulations. In the current implementation, prime activation was applied at the word level, as in Dell's (1986) original implementation. We reasoned that this would allow us to simulate high level activation of other concepts. In architectures with no cascading from phonemes to features, or cascading from selected phonemes to features, prime activation cannot be transmitted to the featural level from unselected phonemes. In the architecture with no feedback from phonemes to words and cascading from all phonemes to features, we suggest that prime activation decays rapidly as there is no feedback from phonemes to words to support it. Very little prime activation is therefore transmitted from the phoneme to the feature level. Error generation at the featural level in all of these architectures is therefore random, such that a high proportion of non-contextual errors are

Table 7.7: Binomial analysis to determine which two-stage architectures can generate a phonological similarity effect, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating significant phonological similarity effects by chance.

	Specific model counts				**Prob.**	
	Total	Excluded	Generated contextual errors	Significant phonological similarity effect		
Cascading from all Ws to Ps						
No cascading from Ps to Fs	2916	784	848	40	0.609	
Cascading from selected Ps to Fs	2916	801	832	28	$> .9$	
Cascading from all Ps to Fs	2916	808	822	31	$> .9$	
Feedback from Fs to Ps	5832	2963	1106	309	$< .001$	*
Feedback from Ps to Ws						
No cascading from Ps to Fs	5832	2470	1410	43	$> .9$	
Cascading from selected Ps to Fs	5832	2467	1441	51	$> .9$	
Cascading from all Ps to Fs	5832	2756	1199	248	$< .001$	*
Feedback from Fs to Ps	5832	3165	1055	279	$< .001$	*

Key:
Ws = words, Ps = phonemes, Fs = features
Prob. = probability

generated. For enough contextual errors to be generated for the phonological similarity effect to be detected, error rates must be very high. There are few specific models which demonstrate a sufficiently high error rate, and the models which do fail both the error rate and non-contextuality constraints.

Direct priming of phonemic and subphonemic representations may occur however, because of perseveratory influences from a recently produced sound in tongue twisters for example. In a model where priming was applied at the subphonemic level, it would not be necessary for any prime activation to cascade for contextual errors to be generated more frequently at the featural level. We argue that this would allow all two-stage architectures to exhibit the phonological similarity effect whilst also observing the constraints on error rate and non-contextuality of errors. Future simulations will seek to confirm this claim. To otherwise rule out this possibility, it would be necessary to demonstrate that subphonemic priming does not occur and all priming originates from higher levels.

Finally, we consider which two stage architectures can simultaneously exhibit both the lexical bias and phonological similarity effects. Figure 7.11 shows that there

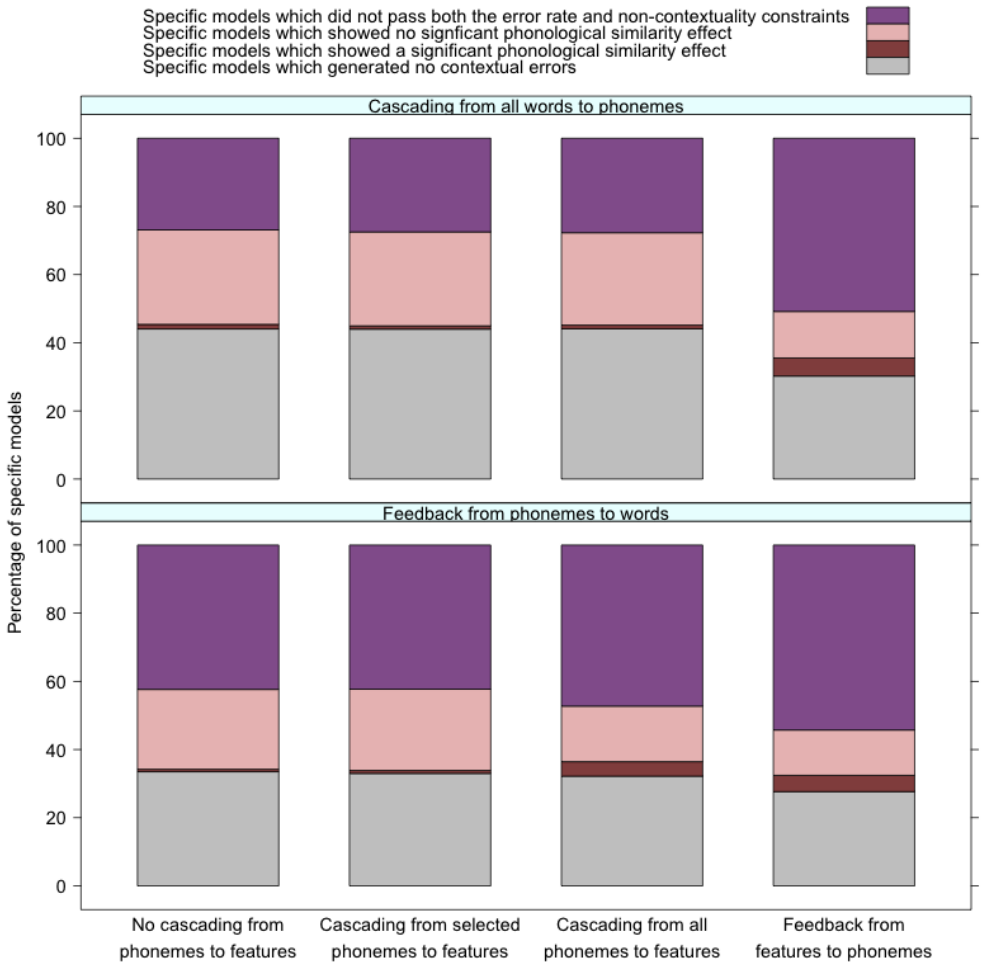


Figure 7.10: The effect of modifying word-to-phoneme and phoneme-to-feature activation flow on exhibition of phonological similarity effects in two-stage models of phonological encoding and subphonemic processing, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

Table 7.8: Binomial analysis to determine which two-stage architectures can generate both a lexical bias and a phonological similarity effect. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating both a significant lexical bias and a significant phonological similarity effect by chance.

	Specific model counts			**Prob.**	
	Total	Generated contextual errors	Significant LB and PS effects		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	1632	9	$> .9$	
Cascading from selected Ps to Fs	2916	1633	4	$> .9$	
Cascading from all Ps to Fs	2916	1630	4	$> .9$	
Feedback from Fs to Ps	5832	4069	62	$> .9$	
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	3880	174	$> .9$	
Cascading from selected Ps to Fs	5832	3908	160	$> .9$	
Cascading from all Ps to Fs	5832	3955	1414	$< .001$	*
Feedback from Fs to Ps	5832	4220	1865	$< .001$	*

Key:
Ws = words, Ps = phonemes, Fs = features
Prob. = probability
LB = lexical bias, PS = phonological similarity

are many more models displaying both effects for the architecture with feedback from phonemes to words and cascading from all phonemes to features, and both the architectures with feedback from features to phonemes, than for the other five architectures. This reflects our findings that due to priming being applied at the word level in the current implementation, only a few models in these other five architectures generate sufficient contextual errors at the featural level for a phonological similarity effect to be detectable. Table 7.8 reports a binomial analysis to determine for which architectures there is evidence of the models' ability to account for both effects simultaneously. Significant results are found only for the architecture with feedback from phonemes to words and cascading from all phonemes to features, and both the architectures with feedback from features to phonemes. We argue that this is largely due to the low numbers of models which generate high enough error rates for the phonological similarity effect to be detected, combined with the reduced power of our binomial analysis for multiple simultaneous effects, as explained in chapter 6. Similar results are found when we exclude specific models which fail either or both of the constraints on error rate or non-contextuality, as shown in figure 7.12 and table 7.9.

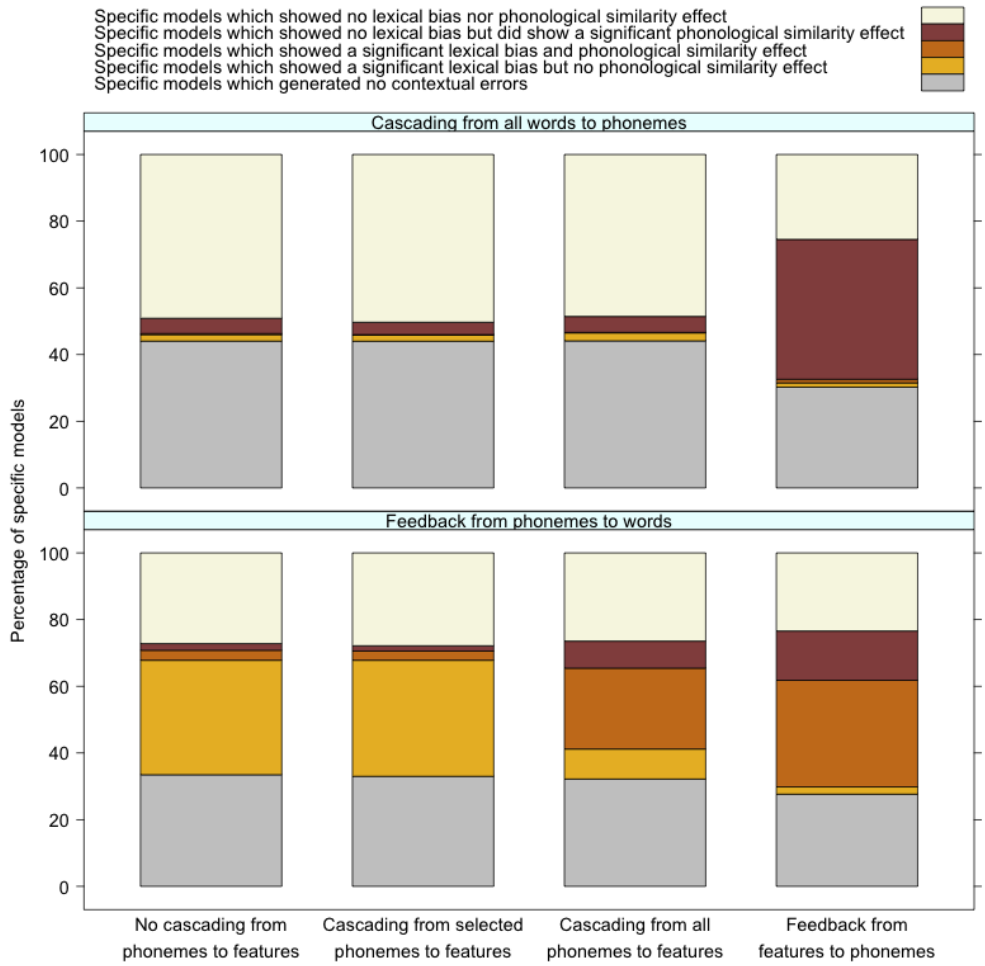


Figure 7.11: The effect of modifying word-to-phoneme and phoneme-to-feature activation flow on exhibition of lexical bias and phonological similarity effects in two-stage models of phonological encoding and subphonemic processing.

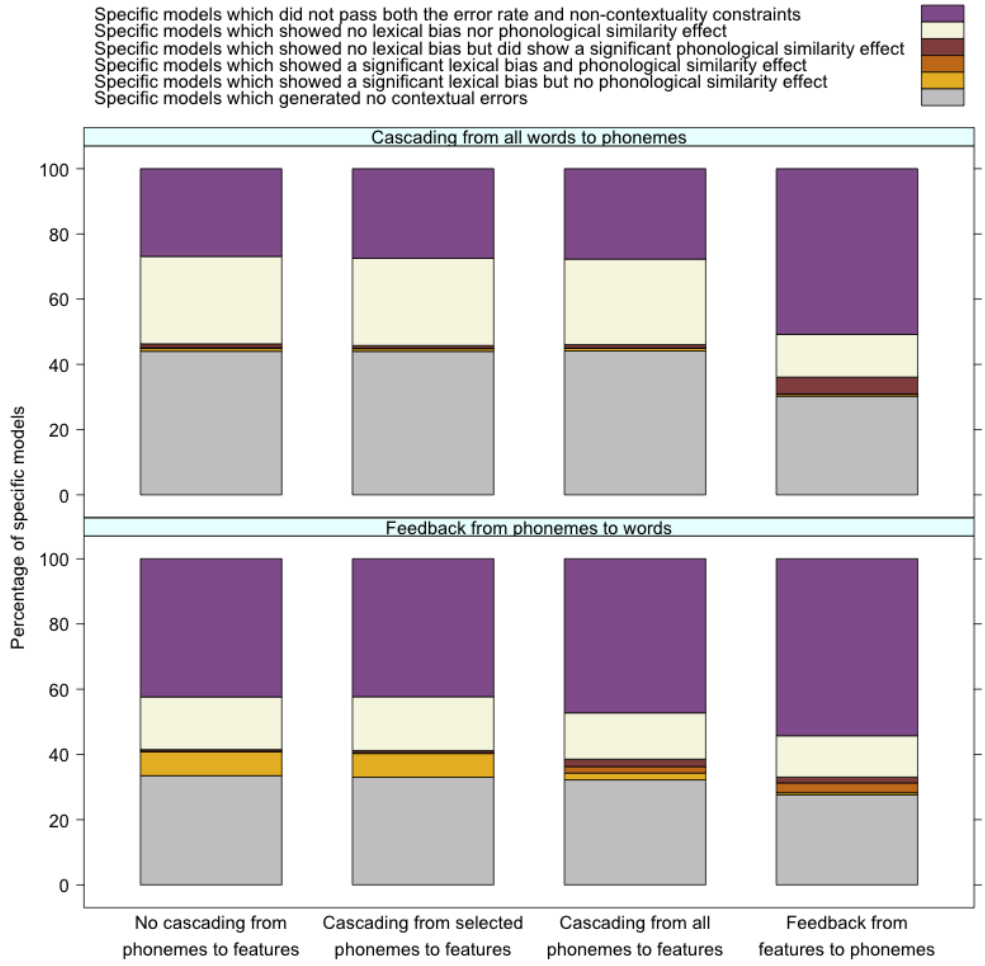


Figure 7.12: The effect of modifying word-to-phoneme and phoneme-to-feature activation flow on exhibition of lexical bias and phonological similarity effects in two-stage models of phonological encoding and subphonemic processing, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

Table 7.9: Binomial analysis to determine which two-stage architectures can generate both a lexical bias and a phonological similarity effect, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating both a significant lexical bias and a significant phonological similarity effect by chance.

	Specific model counts				**Prob.**	
	Total	Excluded	Generated contextual errors	Significant LB and PS effects		
Cascading from all Ws to Ps						
No cascading from Ps to Fs	2916	784	848	5	$> .9$	
Cascading from selected Ps to Fs	2916	801	832	2	$> .9$	
Cascading from all Ps to Fs	2916	808	822	1	$> .9$	
Feedback from Fs to Ps	5832	2963	1106	7	$> .9$	
Feedback from Ps to Ws						
No cascading from Ps to Fs	5832	2470	1410	14	$> .9$	
Cascading from selected Ps to Fs	5832	2467	1441	15	$> .9$	
Cascading from all Ps to Fs	5832	2756	1199	116	$< .001$	*
Feedback from Fs to Ps	5832	3165	1055	175	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

LB = lexical bias, PS = phonological similarity

The effect of spreading activation parameter manipulations on generation of lexical bias and phonological similarity effects

In this section, we examine how parameter manipulations affect the generation of lexical bias and phonological similarity effects.

We showed that there was clear evidence that the architectures with feedback from phonemes to words and either with cascading from all phonemes to features or with feedback from features to phonemes can account for both the lexical bias and phonological similarity effects simultaneously. Here we report the effects of parameter manipulations in the architecture with feedback from phonemes to words and from features to phonemes, but results do not diverge greatly for the architecture with feedback from phonemes to words and cascading from all phonemes to features.

The effects of parameter manipulations on lexical bias generation are similar to those we saw in the one-stage model, as reported in section 6.4. Figure 7.13 and table 7.10 show that high connectivity values, high numbers of steps before selection and high amounts of activation-based noise all increase the number of specific models displaying a significant lexical bias. High decay rates also increase the number of models exhibiting a lexical bias, potentially reflecting the higher numbers of errors generated during the phonological encoding stage when decay rate is high, as well as the reduction in activation of unselected phonemes that high decay rates will cause.

Table 7.11 shows that high connection strengths, high numbers of steps before selection and high levels of activation-based noise increase the probability of specific models showing significant phonological similarity effects. In addition, more phonological similarity effects are observed at high jolt to prime ratios. We suggest that interactive effects are more visible at higher jolt to prime ratios due to the diminished role of the prime on error generation and that this effect outweighs the higher error generation power of lower jolt to prime ratios.

The model of the influence of parameter settings on the combined occurrence of significant lexical bias and phonological similarity effects summarised in table 7.12 fits in with the parameter effects reported for the individual lexical bias and phonological similarity effects. As for both individual models, high connectivity strength, a high number of steps before selection and a high level of activation-based noise make it more likely that models will display both effects. Furthermore, more models

display both effects at high jolt to prime ratios (as for the phonological similarity effect) and at high decay rates (as for the lexical bias effect).

As we have previously argued, we suggest that high connection strengths and high numbers of steps before selection directly support the interactive mechanisms, and high levels of activation-based noise exaggerate the differences in activation levels caused by these mechanisms. However, the status of these parameter settings as settings which lead the models to generate high number of errors and high proportions of non-contextual errors mean that such specific models are likely to get excluded. As figure 7.14 shows, this leaves few specific models which are not excluded by the error constraints but that do generate lexical bias and phonological similarity effects.

We also investigate which parameter settings lead to the generation of phonological similarity effects in the five architectures where prime activation has little effect on featural activation levels (all four architectures with either no cascading from phonemes or cascading from selected phonemes only, and the architecture with no feedback from phonemes to words and cascading from all phonemes). We cannot present a logistic regression analysis for all five architectures together as feedback connection strength is manipulated in two architectures but not in the other three. However, we present the results of parameter manipulations in the two architectures with no feedback from phonemes to words and either no cascading from phonemes or cascading from selected phonemes only, which are representative.

Table 7.13 shows that models with low forward connection strength, high levels of decay, high numbers of steps before selection and high levels of intrinsic noise are all more likely to demonstrate phonological similarity effects. In these models, the signal activation will be weakly transmitted and will decay greatly before feature selection. This will lead to a high number of errors being generated, and consequently a high enough number of contextual errors for the phonological similarity effect to be detected.

Models with high jolt to prime ratios are also more likely to display phonological similarity effects. This reflects the fact that a low prime increases the number of contextual errors generated at the phoneme level, but not the feature level. Contextual errors at the phoneme level will not display a phonological similarity effect in these architectures, whereas contextual errors generated at the featural level will. The phonological similarity effect is therefore diminished at lower jolt to prime ratio as the higher number of errors at the phoneme level weakens the effect.

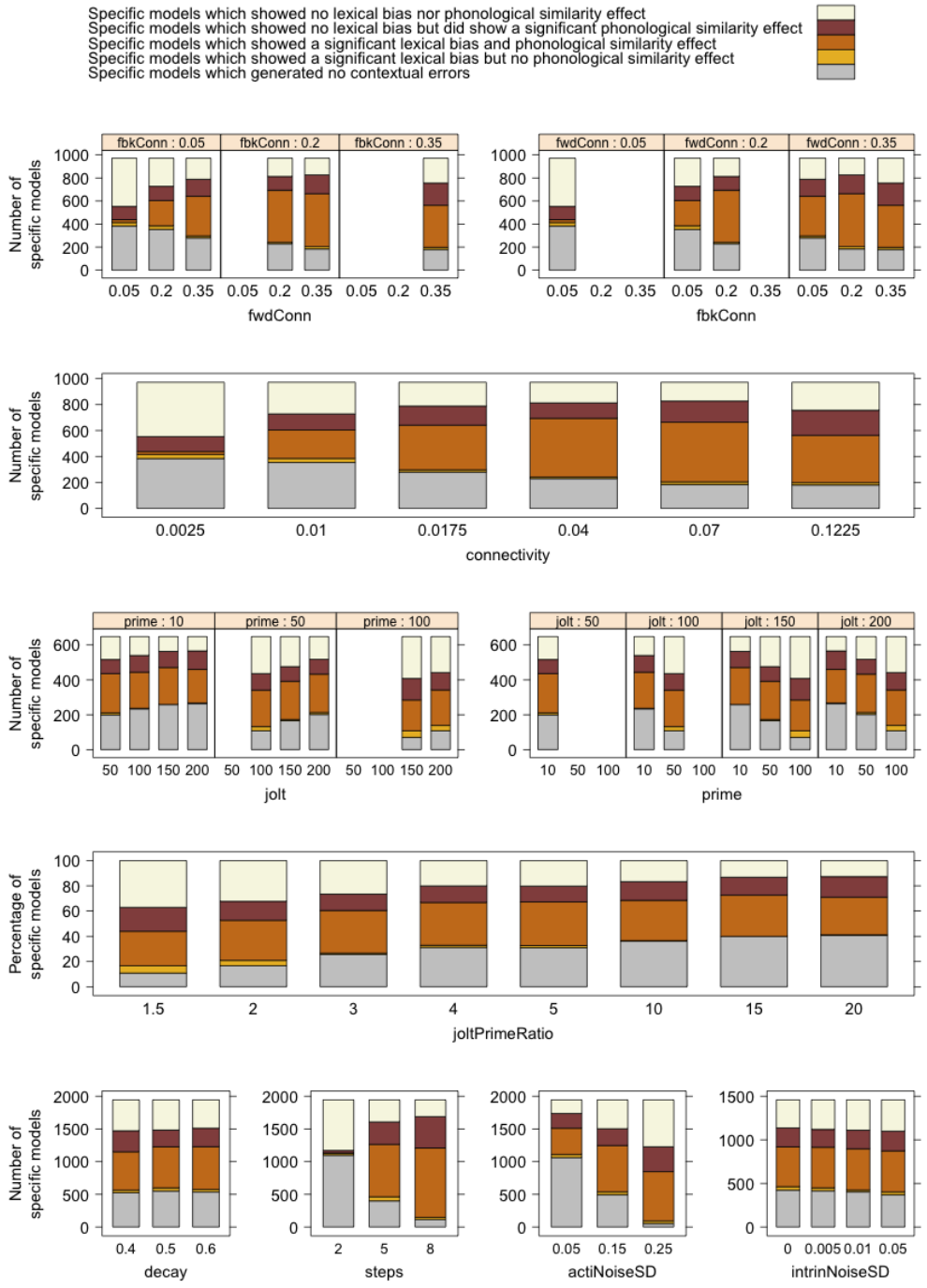


Figure 7.13: The effect of changing parameter settings on exhibition of lexical bias and phonological similarity effects in two-stage phonological encoding models with phoneme-to-word and feature-to-phoneme feedback.

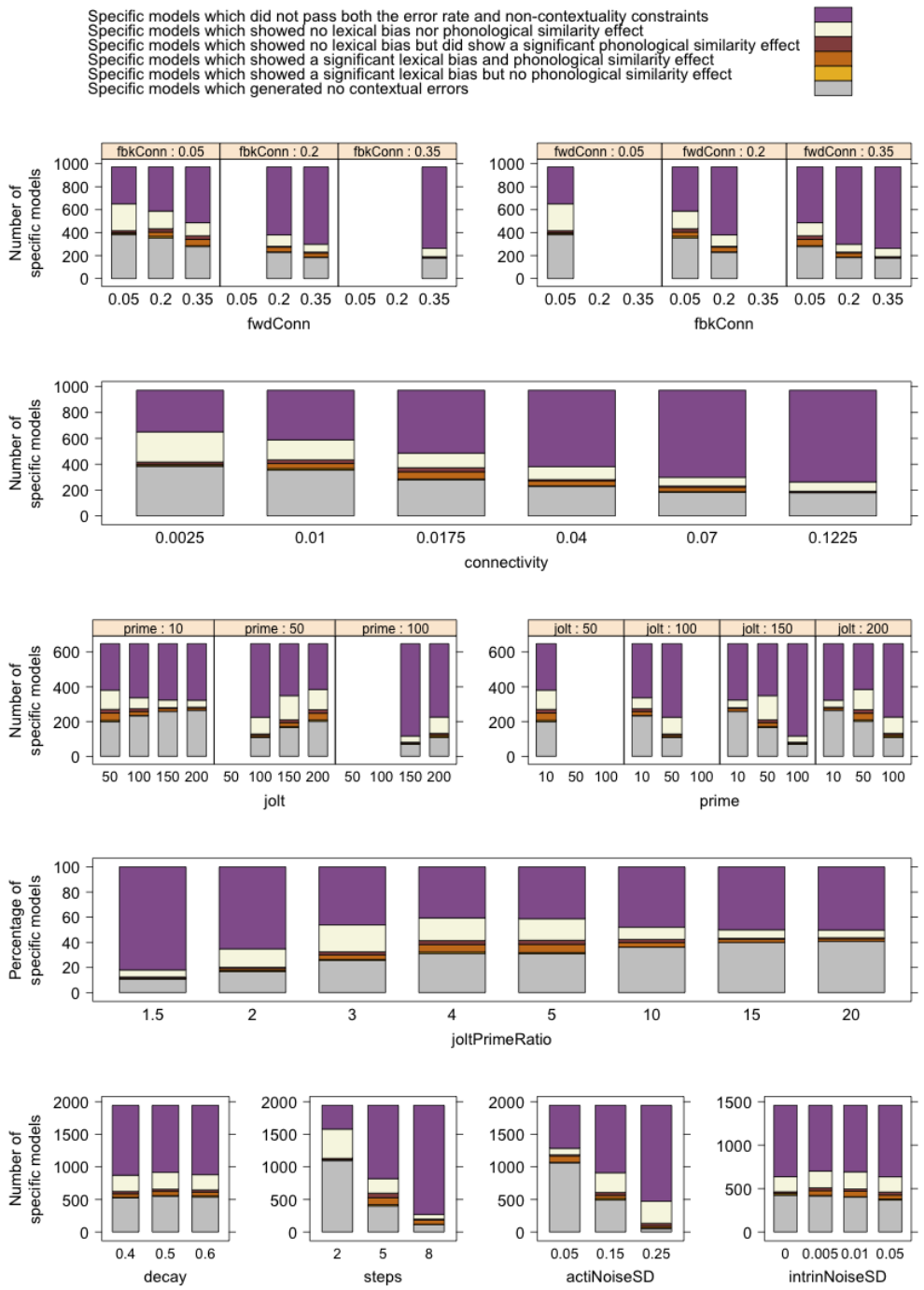


Figure 7.14: The effect of changing parameter settings on exhibition of lexical bias and phonological similarity effects in two-stage phonological encoding models with feature-to-phoneme and phoneme-to-word feedback, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

Table 7.10: Results of logistic regression model analyses using parameter values to predict the occurrence of lexical bias effects, for all two-stage models with phoneme-to-word and feature-to-phoneme feedback connectivity which generated at least one contextual error. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	+	10.3	109	< .001	*
joltPrimeRatio	+	1.3	2	0.21	
decay	+	2.8	8	0.005	*
steps	+	22.8	605	< .001	*
actiNoiseSD	+	3.2	10	0.001	*
intrinNoiseSD	−	0.8	1	0.445	

Table 7.11: Results of logistic regression model analyses using parameter values to predict the occurrence of phonological similarity effects, for all two-stage models with phoneme-to-word and feature-to-phoneme feedback connectivity which generated at least one contextual error. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	+	15.4	273	< .001	*
joltPrimeRatio	+	6.3	42	< .001	*
decay	+	1.6	3	0.113	
steps	+	30.7	1491	< .001	*
actiNoiseSD	+	8.6	76	< .001	*
intrinNoiseSD	−	0.2	0	0.82	

Table 7.12: Results of logistic regression model analyses using parameter values to predict the occurrence of lexical bias and phonological similarity effects, for all two-stage models with both phoneme-to-word and feature-to-phoneme feedback connectivity which generated at least one contextual error. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	+	11.7	142	< .001	*
joltPrimeRatio	+	3.4	12	0.001	*
decay	+	2.9	8	0.004	*
steps	+	24.1	704	< .001	*
actiNoiseSD	+	5.3	28	< .001	*
intrinNoiseSD	−	0.5	0	0.6	

Table 7.13: Results of logistic regression model analyses using parameter values to predict the occurrence of phonological similarity effects, for all two-stage models which generated at least one contextual error with no phoneme-to-word feedback connectivity, and no feature-to-phoneme feedback connectivity (specific models using any of the other three phoneme-to-feature connectivity options are included in the regression). Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
fwdConn	−	2.1	4	0.034	*
joltPrimeRatio	+	3.7	13	< .001	*
decay	+	7.1	54	< .001	*
steps	+	8.9	90	< .001	*
actiNoiseSD	+	0.5	0	0.633	
intrinNoiseSD	+	9.7	92	< .001	*

7.3.3 Conclusions

In this section, we showed that all two-stage architectures with feedback from phonemes to words exhibit the lexical bias effect. More significantly, we demonstrated that all two-stage architectures exhibit the phonological similarity effect. Unlike for the one-stage model, feedback from features to phonemes is not required, and even a model with no cascading from phonemes to features can account for this result because misselection of one feature is more likely than misselection of two.

However, as a result of the decision to apply priming at the word level in this implementation, in the architectures with no cascading from phonemes to features or cascading from selected phonemes only, as well as the architecture with no feedback from phonemes to words and cascading from all phonemes, error generation at the feature level is barely affected or is not affected at all by the prime activation. High proportions of non-contextual errors are therefore generated at the feature level, such that a very high error rate is required for sufficient contextual errors to be produced for the phonological similarity effect to be detected. Models which generate sufficient errors are therefore ruled out by the constraints on error rate and non-contextuality of errors. Future research will verify that applying priming at the featural level removes this problem. Even with priming at the word level however, the architecture with feedback from phonemes to words and cascading from all phonemes exhibits the phonological similarity effect without feedback from features to phonemes.

Lastly, we showed that in architectures with feedback from features to phonemes, or feedback from phonemes to words and cascading from all phonemes, parameters which support activation flow through feedback loops are of core importance in determining whether lexical bias and phonological similarity effects are generated, as was found for the one-stage model. We also showed that in architectures where the prime activation has little effect on error generation at the feature level, models rely on very weak signals and very high levels of decay and noise in order to generate sufficiently high numbers of errors for the phonological similarity effect on contextual errors to be observable.

7.4 Goldrick and Blumstein's (2006) acoustic evidence of traces of intended phonemes on errors

In this section, we build on the work presented in this thesis so far by presenting a simulation of an experiment in which word production is measured instrumentally. Whilst Goldrick and Blumstein (2006) claimed that their evidence of traces of intended phonemes on errors demonstrated cascading from all phonemes to features, we argued that traces could be generated by two other mechanisms (errors at the featural level and weakened activation of unintended but selected phonemes) such that any architecture should be able to account for these results.

7.4.1 Simulation methodology

To evaluate the ability of different models of activation flow between phonemes and features to account for Goldrick and Blumstein's (2006) evidence, we ran a new set of simulations, the details of which are outlined here.

Model configuration

Again, we examined the behaviour of all 37,908 two-stage models.

Model task and lexicon

Goldrick and Blumstein's (2006) results concerned the production of voiced and voiceless stop consonants. To keep the simulation as simple as possible to facilitate examination of the model's behaviour, we used an abstraction of this task and focused on the network's behaviour on productions of the words *"gap"* and *"cap"*. The influence of a competing onset with contrasting voicing is crucial to the setup

of Goldrick and Blumstein's (2006) experiment however. For productions of *"gap"*, the competitor *"cap"* was therefore primed as if it was the upcoming word, and for productions of *"cap"*, the competitor *"gap"* was primed. Each specific model had to attempt the production of *"gap"* 5,000 times and *"cap"* 5,000 times, resulting in a total of 10,000 trials per specific model.

The same 50 word model lexicon was used as in Experiments 1 and 2. As in the experiments detailed in chapter 6 and section 7.3, the lexicon contained words other than the two target words in order to simulate opportunities for non-contextual errors, and to permit the simulations to take some account of the effect of lexicon structure on word production behaviour.

Model output interpretation

In Goldrick and Blumstein's (2006) experiment, the transcriber identified voicing errors by ear. In this experiment therefore, as in others outlined in this chapter, the identity of the onset produced was determined by examining which onset features were most activated at the end of subphonemic processing. Productions were then classified into correct productions, contextual errors and non-contextual errors.

The core analysis then focused on the correct and contextual error productions. For these productions, the simulated VOT measure was calculated from the activation of the voiceless and voiced features, as outlined in chapter 3. Statistical tests to determine whether traces were present were carried out on each specific model. One t-test compared the VOT of all correct productions of /k/ (/k/ → [k]) to all unintended productions of /k/ (/g/ → [k]) to determine whether there are traces of intended voiced productions on voiceless productions. Another t-test compared the VOT of all correct productions of /g/ (/g/ → [g]) to all unintended productions of /g/ (/k/ → [g]) to determine whether there are traces of intended voiceless productions on voiced productions.

However, to aid understanding of the model's behaviour, the identity of the phoneme selected at each trial was also recorded. This permitted comparisons of the VOT of productions where the intended phoneme was selected at the phoneme level (e.g., /k/→/k/) to the VOT of productions where the competing phoneme was selected at the phoneme level (e.g., /g/→/k/). Any traces visible on such comparisons would have to be due to processes affecting phoneme selection, rather than just feature selection. In other words, traces arising on feature errors would not be detectable in these comparisons.

Finally, the activation level of a phoneme when it was selected was also recorded. This was to allow a comparison of the activation level of intentionally selected phonemes (e.g., /k/→/k/) and unintentionally selected phonemes (e.g., /g/→/k/) to confirm our hypothesis that intentionally selected phonemes are more strongly activated than unintentionally selected phonemes.

T-tests were only carried out when there were at least two relevant correct productions and two relevant contextual errors produced, otherwise the specific model was marked as not having enough data for the analysis.

7.4.2 Simulation results

This section presents the results of the simulations. Firstly, as these analyses consider /k/ → [g] and /g/ → [k] errors in particular, the effect of altering word-to-phoneme and phoneme-to-feature activation flow assumptions on these error rates is outlined. Secondly, we report on which architectures can and cannot capture Goldrick and Blumstein's (2006) trace evidence, and to what extent this result is affected by excluding specific models which fail either of the constraints on error rate and non-contextuality of errors, as introduced in chapter 4. Thirdly, the role of the spreading activation parameter settings in these results is examined. Finally, we consider whether any parameter and connectivity setting of the model allows it to account for the transcribed lexical bias effect, the transcribed phonological similarity effect, and Goldrick and Blumstein's (2006) evidence simultaneously, again considering how the error rate and non-contextuality constraints affect these conclusions.

/k/ → [g] and /g/ → [k] error rates

The current analysis focuses specifically on /k/ → [g] and /g/ → [k] errors. Figure 7.15 demonstrates the effect of manipulating activation flow assumptions on these error rates. For extra clarity, the median error rates for each architecture are provided in table 7.14.

Some result patterns fit in with our previous analyses of the effect of manipulating activation flow on error rate in section 7.2. For example, there is a general tendency for error rate to increase when feedback from features to phonemes is added to the model; an overall higher error rate when feedback from phonemes to words is present; and for architectures with feedback from phonemes to words, higher error

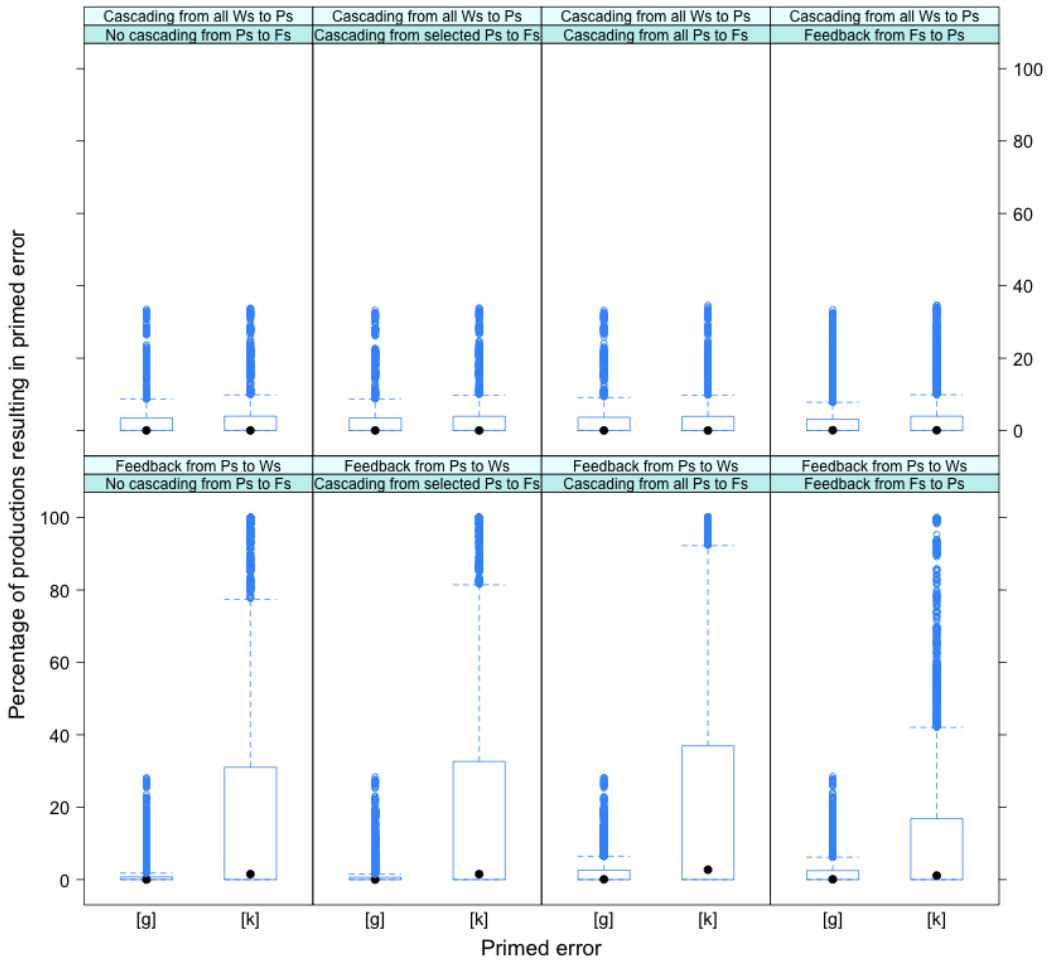


Figure 7.15: The effect of modifying activation flow on the rate of primed /g/ and /k/ errors in two-stage models.
Key: Ws = words, Ps = phonemes, Fs = features

Table 7.14: Median rates of /k/ → [g] errors as a percentage of all attempted /k/ productions, and /g/ → [k] errors as a percentage of all attempted /g/ productions.

	Median rate of /k/ → [g] errors (%)	**Median rate of /g/ → [k] errors (%)**
Cascading from all Ws to Ps		
No cascading from Ps to Fs	0.04	0.04
Cascading from selected Ps to Fs	0.04	0.04
Cascading from all Ps to Fs	0.04	0.04
Feedback from Fs to Ps	0.10	0.10
Feedback from Ps to Ws		
No cascading from Ps to Fs	0.00	1.50
Cascading from selected Ps to Fs	0.00	1.50
Cascading from all Ps to Fs	0.06	2.71
Feedback from Fs to Ps	0.08	1.10

Key:
Ws = words, Ps = phonemes, Fs = features

rates for the architecture with cascading from all phonemes in comparison to the architecture with cascading from selected phonemes only.

Other results are particular to this analysis however. Firstly, the inclusion of phoneme-to-word feedback clearly increases the rate of /g/ → [k] errors, whilst reducing the rate of /k/ → [g] errors. This can easily be explained as being due to the far higher frequency of /k/ as an onset phoneme in our model lexicon, where /k/ is the onset phoneme of 7 words, in comparison to /g/, which is the onset phoneme of 2 words. As detailed in section 4.2.2, words in this lexicon were selected largely at random from the BEEP dictionary of English words. Correspondingly, this difference in phoneme frequency reflects the underlying frequencies of these phonemes in the English language, where Shattuck-Hufnagel and Klatt (1979) report that /k/ is used between 1.75 and 4 times more often than /g/ in English conversation.

The higher frequency of /k/ as an onset phoneme means that when phoneme-to-word feedback is present, the activation of /k/ is boosted in comparison to a less frequent onset. As also explained in section 6.2.2, when feedback from phonemes to words is present, any activation which an onset phoneme possesses is transmitted via the feedback connections to words in which the onset phoneme participates. Each word receives from the onset phoneme an amount of activation equal to the activation level of the onset phoneme multiplied by the strength of the feedback connection. There is no notion of the activation being shared between words; i.e., the amount of activation received from an onset phoneme by a word does not decrease simply because more words are connected to the onset phoneme. Crucially, on the following spreading activation step, all of the connected words send activation back to the onset phoneme. Activation sent back to the onset phoneme therefore increases as the number of connected words increases. As a result, the onset phoneme /k/ tends to receive more activation than /g/, such that it happens more frequently that /k/ is more activated than /g/ than vice versa. This behaviour fits in with results reported by Levitt and Healy (1985), which showed that attempted productions of less frequent phonemes are more prone to errors, and more frequent phonemes are more likely to intrude.

Secondly, we note that in architectures with feedback from phonemes to words, the rate of /g/ → [k] errors decreases when feedback from features to phonemes is added to a model. We suggest this is because there are more voiced consonants than voiceless consonants in English, as is reflected in this lexicon, where 9 of the included onset consonants are voiced and 7 are voiceless. Feedback from features to phonemes will therefore cause the voiced feature to receive more activation than

the voiceless feature, increasing the probability of [voiceless] → [voiced] errors and reducing the probability of [voiced] → [voiceless] errors.

Traces on voiced outcome productions and voiceless outcome productions

Goldrick and Blumstein (2006) found that participants in their experiments generated traces on both voiced and voiceless productions. Here, we begin by examining which models can account for voiced traces (where /g/ → [g] productions are compared to /k/ → [g] productions) and which can account for voiceless traces (where /k/ → [k] productions are compared to /g/ → [k] productions). The following section considers which models can account for both types of traces simultaneously.

Figure 7.16 shows how many specific models generated voiced traces and how many generated voiceless traces for each architecture. Table 7.15 presents a binomial analysis to determine for each architecture whether enough models display traces of intended voiceless productions on voiced productions to accept that the architecture can account for this evidence, and table 7.16 presents a similar analysis for traces of intended voiced productions on voiceless productions.

We begin by considering results in architectures with no feedback from phonemes to words. For both voiced and voiceless traces, the graph and statistics suggest that as we had hypothesised, all architectures can account for this evidence. Furthermore, as activation flow between phonemes and features became more interactive, the number of specific models displaying traces increased. This fits in with our argument that there are progressively more ways that traces can be generated as interactivity increases. The architecture with no cascading from phonemes to features can generate traces but only on featural errors; the architecture with cascading from selected phonemes to features can generate traces for this reason, but also because intentionally selected phonemes are more strongly activated than unintentionally selected phonemes; and the architecture with cascading from all phonemes to features can generate traces for both these reasons, but also because when a phoneme is selected in error, activation cascades from the intended but unselected phoneme. An increase in the number of mechanisms by which traces can be generated may result in a stronger trace which is more likely to be detected, or it may mean that specific models with parameter settings which were not conducive to trace generation by another mechanism can now generate traces. Even more traces are generated when feedback from features to phonemes is present. We argue that this is not due to a new trace generating mechanism, but because more errors are

Table 7.15: Binomial analysis to determine which two-stage architectures exhibit traces of intended voiceless phonemes on voiced productions. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating significant traces by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	Significant traces		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	1460	215	$< .001$	*
Cascading from selected Ps to Fs	2916	1458	418	$< .001$	*
Cascading from all Ps to Fs	2916	1460	593	$< .001$	*
Feedback from Fs to Ps	5832	3077	2051	$< .001$	*
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	2518	357	$< .001$	*
Cascading from selected Ps to Fs	5832	2520	725	$< .001$	*
Cascading from all Ps to Fs	5832	2498	1656	$< .001$	*
Feedback from Fs to Ps	5832	2756	1857	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

Table 7.16: Binomial analysis to determine which two-stage architectures exhibit traces of intended voiced phonemes on voiceless productions. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating significant traces by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	Significant traces		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	1486	215	$< .001$	*
Cascading from selected Ps to Fs	2916	1473	430	$< .001$	*
Cascading from all Ps to Fs	2916	1477	597	$< .001$	*
Feedback from Fs to Ps	5832	3113	2029	$< .001$	*
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	3803	298	$< .001$	*
Cascading from selected Ps to Fs	5832	3793	2427	$< .001$	*
Cascading from all Ps to Fs	5832	3790	3057	$< .001$	*
Feedback from Fs to Ps	5832	3725	2941	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

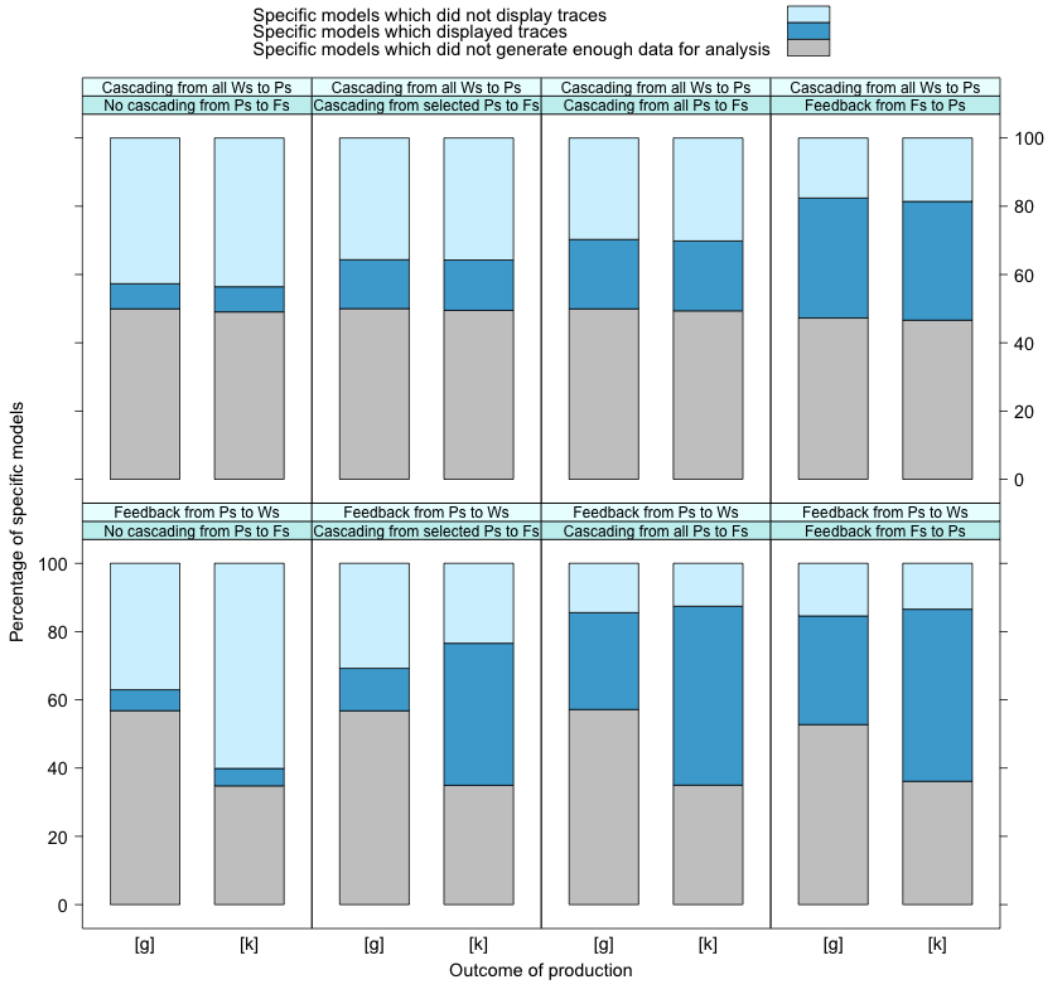


Figure 7.16: The effect of modifying activation flow on trace generation on /k/ and /g/ productions in two-stage models.
Key: Ws = words, Ps = phonemes, Fs = features

generated when feedback from features to phonemes is present such that more data is available for the analyses, and smaller effects become easier to detect.

To further increase our understanding of how these different architectures are generating traces, in figure 7.17, we explicitly mark specific models in which no traces are generated at phoneme selection. Whether traces are generated at phoneme selection or not can be established by comparing the VOT of productions where the intended phoneme was selected at the phoneme level to the VOT of productions where the competing phoneme was selected at the phoneme level. In other words, the identity of the phoneme selected at the phoneme level is considered instead of the identity of the phoneme selected at the feature level. This analysis is therefore not sensitive to traces generated at feature selection and can only detect traces generated at phoneme selection. Similarly, in table 7.19, we report as a percentage of all

the models tested for that architecture, the total number of models which displayed traces; the number of models which displayed traces both when productions were classified at the featural level and when productions were classified at the phoneme level (such that some of the traces must have originated in phonological encoding); and the number of models which only displayed traces when productions were classified at the featural level, but did not display traces when productions were classified at the phoneme level (implying that traces were not reliably generated at the phonological encoding stage).

Figure 7.17 and table 7.19 show that a particularly high proportion of the models with no cascading from phonemes to features are generating traces at the feature level only, in line with our hypothesis about trace generation in this model. Furthermore, tables 7.17 and 7.18 demonstrate that for both voiced and voiceless productions, the number of models which display significant traces when productions are classified at the phoneme level as well as displaying significant traces when productions are classified at the featural level is not bigger than chance would predict for the architecture with no cascading from phonemes to features, although it is for all of the other architectures. We note again that as a model must demonstrate two significant effects to pass this test, the per specific model chance of a Type I error such that at least one of these effects was significant by chance is $1 - (0.95 \times 0.95)$, and our calculations take this into account.

When feedback from phonemes to words is present, there are far more voiceless outcome traces than voiced outcome traces. This is in line with our earlier observation that when the model contains phoneme-to-word feedback, many more /k/ outcome errors than /g/ outcome errors occur, so that there is more data available for voiceless trace analyses.

Results are otherwise largely similar, with a couple of interesting extra points. Firstly, the increase in the number of specific models generating traces from the architecture with cascading from selected phonemes only to the architecture with cascading from all phonemes is bigger when phoneme-to-word feedback is present, as can be verified by examining table 7.19. In section 7.2, we argued that the architecture with feedback from phonemes to words and cascading from phonemes to features generates a lot of errors due to noise which originates in the phoneme-to-word feedback loop and cascades to the featural level, causing featural errors. This noise does not cascade in models with cascading from selected phonemes only because the restricted activation flow makes this impossible, and in models with no feedback from phonemes to words, the noise does not build up in the first place.

Table 7.17: Binomial analysis to determine which two-stage architectures display traces on voiced productions both when productions are classified at the featural level and when productions are classified at the phoneme level, such that some of the traces must have originated in phonological encoding. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating significant traces by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	Significant traces		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	1460	89	$> .9$	
Cascading from selected Ps to Fs	2916	1458	294	$< .001$	*
Cascading from all Ps to Fs	2916	1460	442	$< .001$	*
Feedback from Fs to Ps	5832	3077	1749	$< .001$	*
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	2518	181	$> .9$	
Cascading from selected Ps to Fs	5832	2520	632	$< .001$	*
Cascading from all Ps to Fs	5832	2498	795	$< .001$	*
Feedback from Fs to Ps	5832	2756	1123	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

Table 7.18: Binomial analysis to determine which two-stage architectures display traces on voiceless productions both when productions are classified at the featural level and when productions are classified at the phoneme level, such that some of the traces must have originated in phonological encoding. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating significant traces by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	Significant traces		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	1486	95	$> .9$	
Cascading from selected Ps to Fs	2916	1473	321	$< .001$	*
Cascading from all Ps to Fs	2916	1477	461	$< .001$	*
Feedback from Fs to Ps	5832	3113	1745	$< .001$	*
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	3803	173	$> .9$	
Cascading from selected Ps to Fs	5832	3793	2331	$< .001$	*
Cascading from all Ps to Fs	5832	3790	2480	$< .001$	*
Feedback from Fs to Ps	5832	3725	2559	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

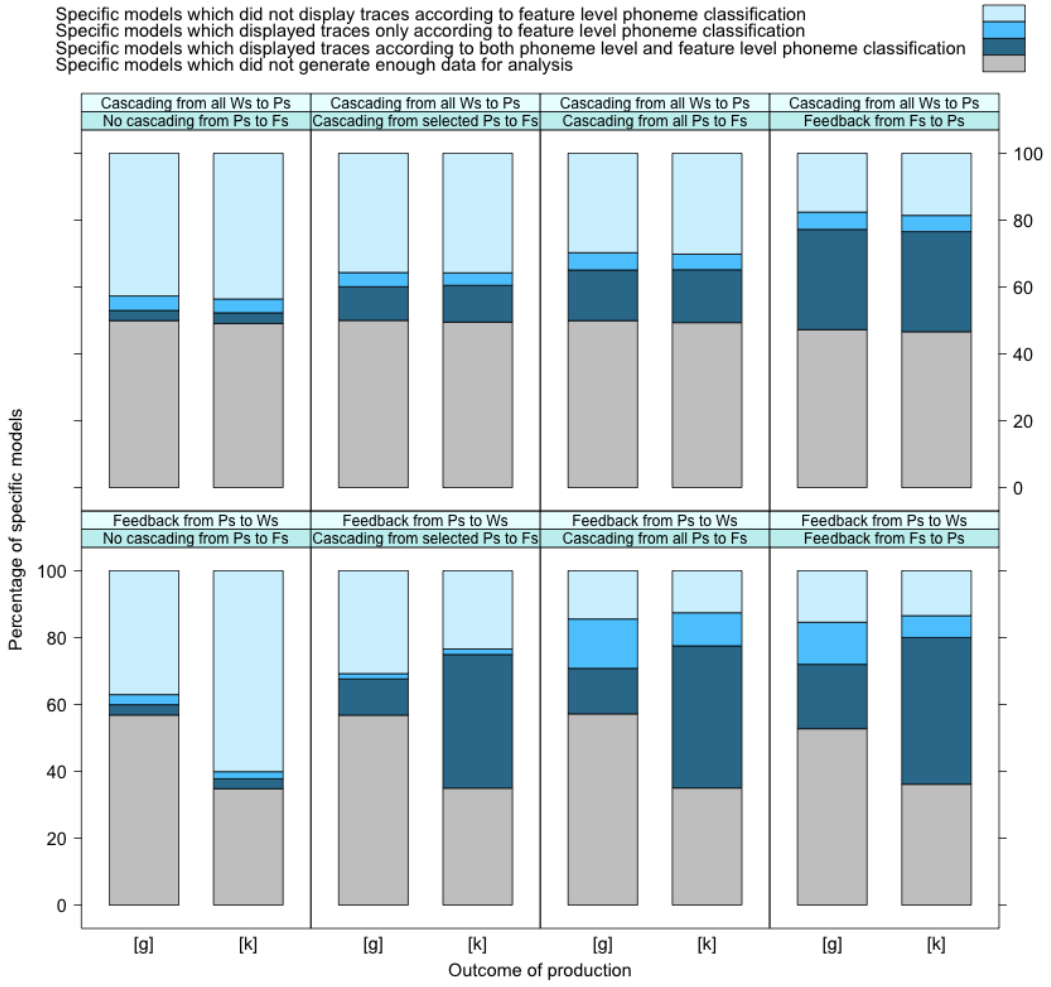


Figure 7.17: The effect of modifying activation flow on trace generation on /k/ and /g/ productions in two-stage models, considering whether traces originated in phonological encoding or not.
Key: Ws = words, Ps = phonemes, Fs = features

Table 7.19: The effect of modifying activation flow on trace generation on /k/ and /g/ productions in two-stage models, considering whether traces originated in phonological encoding or not. Numbers represent the percentage of models tested for that architecture which displayed the specified type of traces.

	Traces on /g/ productions			**Traces on /k/ productions**		
	Feature traces	*Phoneme and feature traces*	*Feature traces only*	Feature traces	*Phoneme and feature traces*	*Feature traces only*
Cascading from all Ws to Ps						
No cascading from Ps to Fs	7.4	*3.1*	*4.3*	7.4	*3.3*	*4.1*
Cascading from selected Ps to Fs	14.3	*10.1*	*4.3*	14.7	*11.0*	*3.7*
Cascading from all Ps to Fs	20.3	*15.2*	*5.2*	20.5	*15.8*	*4.7*
Feedback from Fs to Ps	35.2	*30.0*	*5.2*	34.8	*29.9*	*4.9*
Feedback from Ps to Ws						
No cascading from Ps to Fs	6.1	*3.1*	*3.0*	5.1	*3.0*	*2.1*
Cascading from selected Ps to Fs	12.4	*10.8*	*1.6*	41.6	*40.0*	*1.6*
Cascading from all Ps to Fs	28.4	*13.6*	*14.8*	52.4	*42.5*	*9.9*
Feedback from Fs to Ps	31.8	*19.3*	*12.6*	50.4	*43.9*	*6.6*

Key:

Ws = words, Ps = phonemes, Fs = features

Feature traces = traces according to feature level phoneme classification

Phoneme and feature traces = traces according to both phoneme level and feature level phoneme classification

Feature traces only = traces only according to feature level phoneme classification

It would follow that a number of these extra models showing specific traces in the architecture with cascading from all phonemes may be generating traces due to errors at the featural level. Figure 7.17 and table 7.19 confirm this suggestion.

The story is different for traces generated at phonological encoding, however. In architectures without feedback from phonemes to words, allowing cascading from all phonemes increases the number of specific models displaying traces generated at phonological encoding, in comparison to when activation cascades from selected phonemes only. This is because of the extra mechanism for trace generation which becomes available (activation cascading from the intended but unselected phoneme). However, in architectures with feedback from phonemes to words, this increase is smaller. We argue that the same noise that cascades from the phoneme-to-word feedback loop and causes featural error production and trace generation, also reduces the number of traces stemming from phoneme selection which can be detected. This results from noise distorting the patterns of activation which are transmitted from the phoneme to the feature level, and adding more variance to the VOTs recorded in each condition.

A similar argument can be made about behaviour in architectures with feedback from phonemes to words, and from features to phonemes. As noted before, when there is no feedback from phonemes to words, adding feedback from features to phonemes increases the number of specific models which display traces. This was attributed to the feedback creating noise which increased the error rate and boosted the amount of data available to the analysis. However, as can be seen in figure 7.16 when feedback from phonemes to words is present, the increase in the number of specific models displaying traces caused by adding feedback from features to phonemes is much smaller. In fact, the overall number of specific models displaying traces on [k] productions is smaller when feedback from features to phonemes is added. Again, we suggest that whilst a certain level of noise boosts the number of models displaying traces because more data is available, noise above this level distorts the patterns of activation created at phoneme selection too far for traces to be detected, and reduces the power of the statistical test by increasing the variance of the VOTs in each condition.

However, as for our investigation of the transcribed phonological similarity effect. Figure 7.18 shows that in architectures where prime activation from the word level does not effectively reach the featural level, specific models which rely on error generation at the featural level to exhibit the phonological similarity effect are ruled out when the constraints on error rate or non-contextuality are applied. This is particularly problematic for the architecture with no cascading from phonemes to features, as it cannot generate traces on errors at the phoneme level. Table 7.20 shows that for traces on voiced productions, the number of models showing significant traces is not greater than would be predicted by chance when there is no cascading from phonemes and no feedback from phonemes to words is present. Table 7.20 also shows that for traces on voiceless productions, the number of models showing significant traces is not greater than would be predicted by chance when there is no cascading from phonemes and feedback from phonemes to words is present. (In this case, we cannot see an obvious reason why the presence of feedback from phonemes to words would have an effect on these results, and explain the different results for different phoneme-to-word feedback settings as being due to a combination of a shrinking set of models which can account for this evidence and random noise.) We note again that an implementation in which priming was applied at the featural level would be unlikely to experience these problems. Future research will seek to verify this claim.

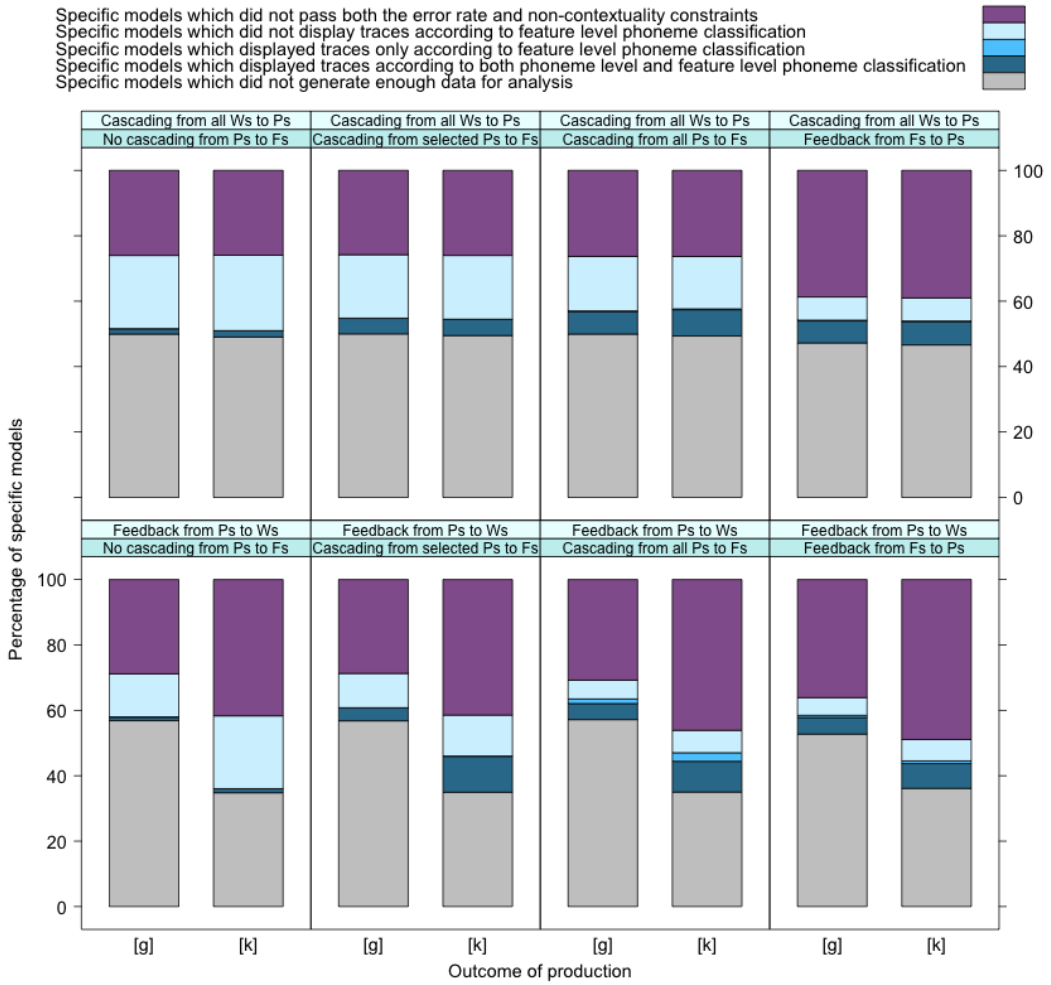


Figure 7.18: The effect of modifying activation flow on trace generation on /k/ and /g/ productions in two-stage models, considering whether traces originated in phonological encoding or not, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.
Key: Ws = words, Ps = phonemes, Fs = features

Table 7.20: Binomial analysis to determine which two-stage architectures exhibit traces of intended voiceless phonemes on voiced productions, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating significant traces by chance.

	Specific model counts				**Prob.**	
	Total	Excluded	Sufficient data	Significant traces		
Cascading from all Ws to Ps						
No cascading from Ps to Fs	2916	758	702	49	0.009	
Cascading from selected Ps to Fs	2916	752	706	140	$< .001$	*
Cascading from all Ps to Fs	2916	767	693	207	$< .001$	*
Feedback from Fs to Ps	5832	2258	819	406	$< .001$	*
Feedback from Ps to Ws						
No cascading from Ps to Fs	5832	1682	836	69	$< .001$	*
Cascading from selected Ps to Fs	5832	1675	845	233	$< .001$	*
Cascading from all Ps to Fs	5832	1795	703	368	$< .001$	*
Feedback from Fs to Ps	5832	2106	650	334	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

Table 7.21: Binomial analysis to determine which two-stage architectures exhibit traces of intended voiced phonemes on voiceless productions, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating significant traces by chance.

	Specific model counts				**Prob.**	
	Total	Excluded	Sufficient data	Significant traces		
Cascading from all Ws to Ps						
No cascading from Ps to Fs	2916	757	729	57	$< .001$	*
Cascading from selected Ps to Fs	2916	758	715	146	$< .001$	*
Cascading from all Ps to Fs	2916	768	709	242	$< .001$	*
Feedback from Fs to Ps	5832	2275	838	427	$< .001$	*
Feedback from Ps to Ws						
No cascading from Ps to Fs	5832	2431	1372	73	0.268	
Cascading from selected Ps to Fs	5832	2418	1375	645	$< .001$	*
Cascading from all Ps to Fs	5832	2690	1100	701	$< .001$	*
Feedback from Fs to Ps	5832	2853	872	490	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

Accounting for traces on voiced outcome productions and traces on voiceless outcome productions simultaneously

Goldrick and Blumstein's (2006) evidence showed that humans exhibit both traces of intended voiceless productions on unintended voiced productions and traces of intended voiced productions on unintended voiceless productions. In this section, we investigate which models display both types of traces simultaneously.

Figure 7.19 shows that the basic effect of manipulating the activation flow between words and phonemes, and crucially phonemes to features, is the same when we examine which models display both types of traces simultaneously as it was when we examined which models generated voiced and voiceless traces individually. Specifically, the number of models displaying significant traces increases as activation flow between phonemes and features becomes more interactive. Models with no cascading from phonemes which show significant traces nearly all generate these traces at the feature level. Adding phoneme-to-word feedback to the model increases the number of models which display traces. Models with phoneme-to-word feedback and either cascading from all features or even feedback from features to phonemes, display higher numbers of models which generate traces at the feature level than do the other architectures, presumably due to the level of noise in these models. In line with the individual voiced and voiceless trace results, figure 7.20 shows that again, most specific models which only generate traces at the feature level get ruled out by the constraints on error rate and non-contextuality.

However, table 7.22 shows that the number of models with no cascading from phonemes to features which display traces is not greater than chance would predict, even before the constraints on error rate and non-contextuality of errors are applied. It seems sensible to conclude that this result is due to the combination of the little support for feature level contextual errors due to priming being applied at the word level in the current implementation, and the weaker power of the binomial analysis when seeking evidence for multiple effects (as explained in chapter 6).

We also note that when the constraints on error rate and non-contextuality of errors are applied, and no phoneme-to-word feedback is present in the model, the number of models with cascading from selected phonemes which display traces is also not greater than chance would predict. We suggest that this is due to a combination of fewer phoneme errors being generated when no feedback from phonemes to words is present (reducing the amount of data available for analysis), and trace generation

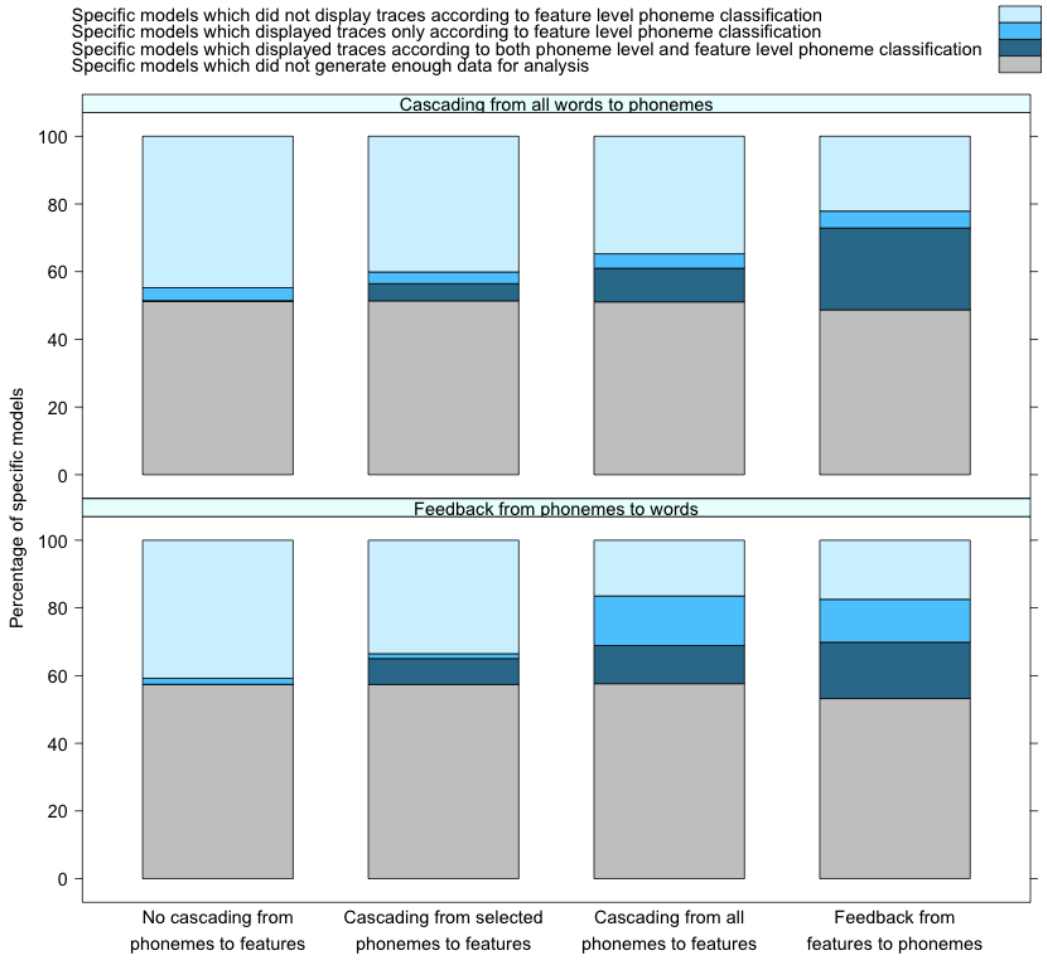


Figure 7.19: The effect of modifying activation flow on models' ability to generate traces on both /k/ and /g/ productions in two-stage models, considering whether traces originated in phonological encoding or not.

relying solely on lower activation levels in unintentionally selected phonemes, such that traces are weaker than in architectures with cascading from all phonemes.

Table 7.22: Binomial analysis to determine which two-stage architectures exhibit traces of intended voiced phonemes on voiceless productions and traces of intended voiceless phonemes on voiced productions. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating voiced or voiceless traces by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	Significant traces		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	1423	118	$> .9$	
Cascading from selected Ps to Fs	2916	1418	249	$< .001$	*
Cascading from all Ps to Fs	2916	1427	414	$< .001$	*
Feedback from Fs to Ps	5832	2990	1700	$< .001$	*
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	2486	107	$> .9$	
Cascading from selected Ps to Fs	5832	2486	531	$< .001$	*
Cascading from all Ps to Fs	5832	2471	1511	$< .001$	*
Feedback from Fs to Ps	5832	2729	1711	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

Table 7.23: Binomial analysis to determine which two-stage architectures exhibit traces of intended voiced phonemes on voiceless productions and traces of intended voiceless phonemes on voiced productions, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating voiced or voiceless traces by chance.

	Specific model counts				**Prob.**	
	Total	Excluded	Sufficient data	Significant traces		
Cascading from all Ws to Ps						
No cascading from Ps to Fs	2916	751	672	5	$> .9$	
Cascading from selected Ps to Fs	2916	747	671	48	$> .9$	
Cascading from all Ps to Fs	2916	761	666	130	$< .001$	*
Feedback from Fs to Ps	5832	2203	787	310	$< .001$	*
Feedback from Ps to Ws						
No cascading from Ps to Fs	5832	1676	810	2	$> .9$	
Cascading from selected Ps to Fs	5832	1665	821	124	$< .001$	*
Cascading from all Ps to Fs	5832	1786	685	300	$< .001$	*
Feedback from Fs to Ps	5832	2095	634	267	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

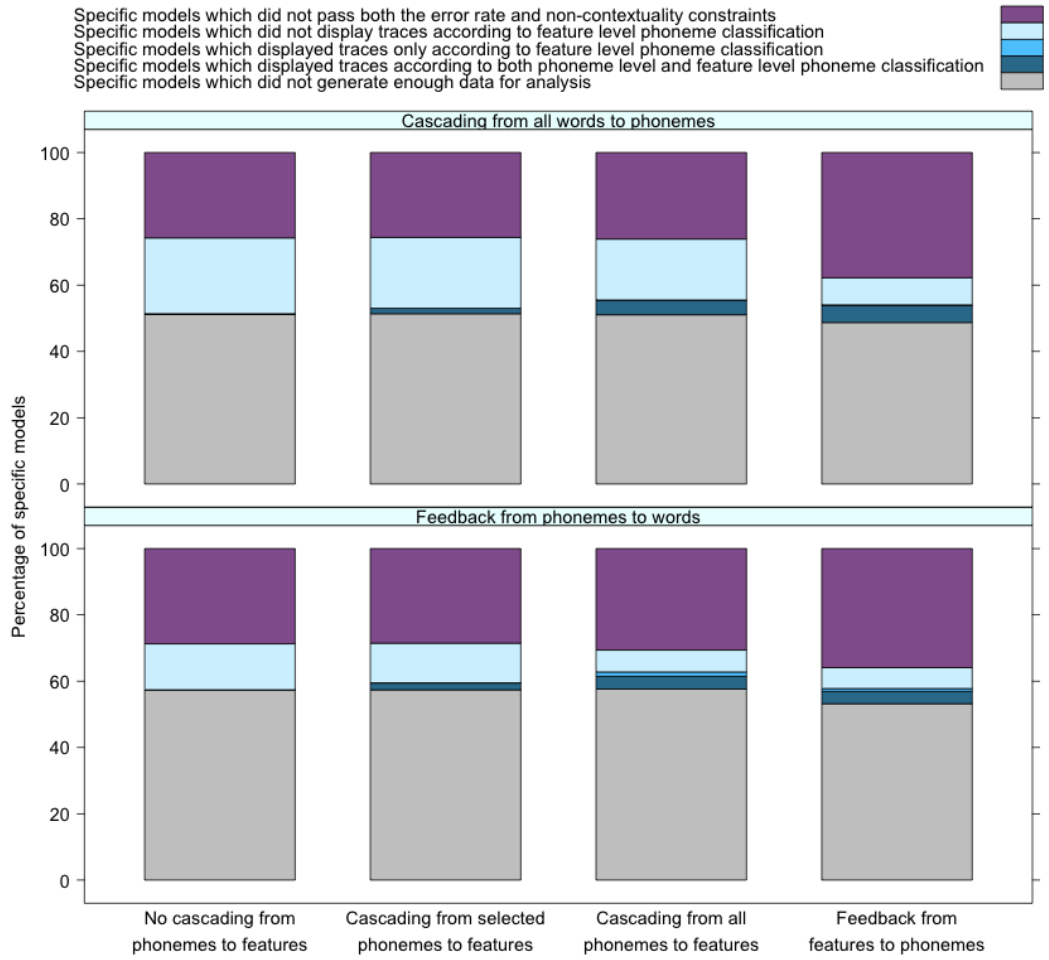


Figure 7.20: The effect of modifying activation flow on models' ability to generate traces on both /k/ and /g/ productions in two-stage models, considering whether traces originated in phonological encoding or not, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

The effect of manipulating spreading activation parameters on trace generation

The previous sections have demonstrated that within every architecture, some specific models display traces. Equally however, for every architecture, there are many specific models which do not generate traces. A simple explanation would be that in these specific models, not enough errors are generated for the effect to be detected. If this is the case, we would expect to find that spreading activation parameter settings which lead to high error rates also lead to high numbers of specific models generating traces. However, it is possible that the parameters are affecting the tendency of the models to generate traces in other ways, and in this section, we examine that possibility further.

Firstly, we verify the general truth of our assumption that intentionally selected phonemes are more strongly activated than unintentionally selected phonemes, and investigate whether any parameter settings cause this not to be the case. Secondly, we look at the effect of spreading activation parameter manipulations on whether traces are generated, taking into consideration whether traces originated at phoneme or feature selection.

Figure 7.21 shows that in nearly every model which generates enough errors for analysis, intended phonemes are more strongly activated at selection than unintended phonemes. The binomial analysis presented in table 7.24 unsurprisingly confirms that for every architecture, the number of models for which this difference is significant is much more than would be predicted by chance. Of course, in the architecture with no cascading from the phoneme level to the feature level, selected phonemes only transmit the jolt activation to the featural level, and so this difference in pre-selection phoneme activation is not evident following phoneme selection and therefore cannot create a trace of the intended phoneme on erroneous productions.

These results both demonstrate the validity our assumption that intended phonemes are more strongly activated at selection than unintended phonemes. Interestingly, the extremely high proportion of models demonstrating this difference also underlines that this effect is generally strong enough for it to be detectable without models generating large numbers of errors for analysis.

Figure 7.21 does highlight however that a few models with feedback from phonemes to words do not display this difference. Figure 7.22 depicts the effect of manipulating spreading activation parameters on whether this difference is detected, for all

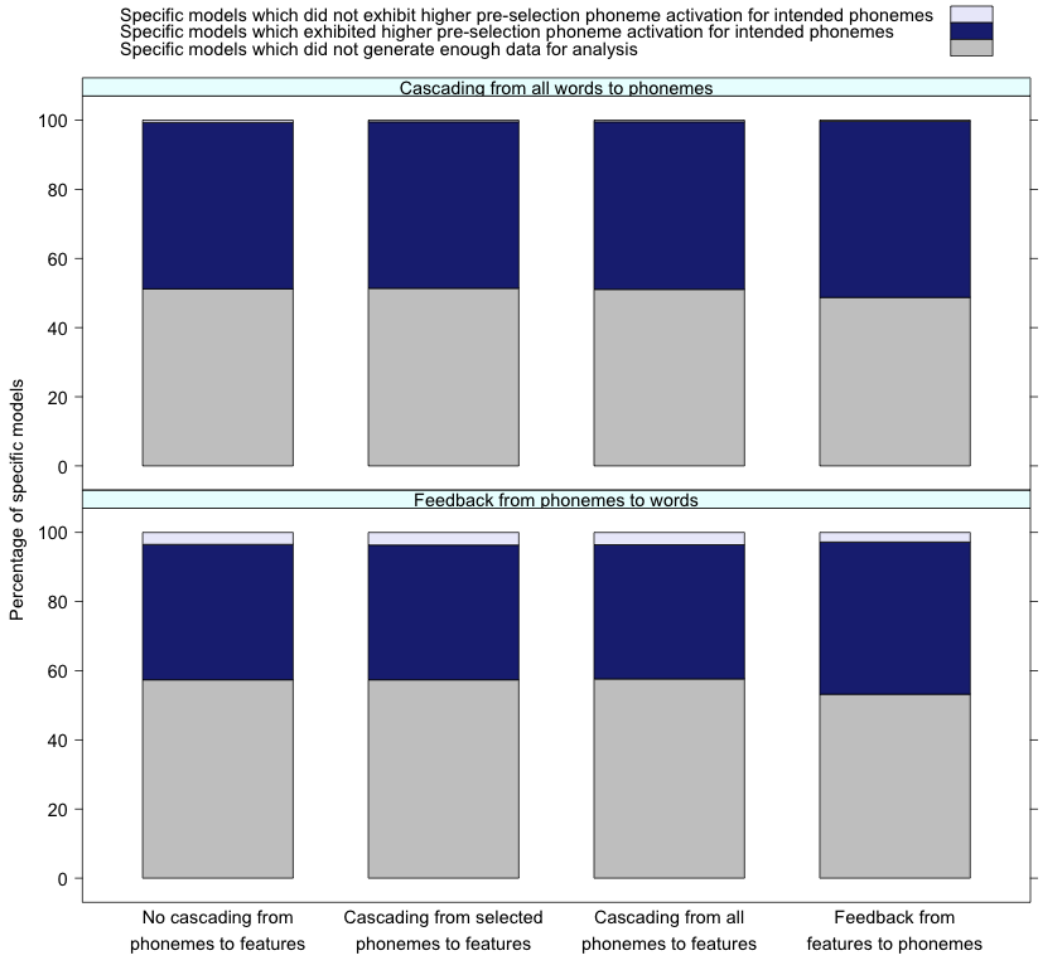


Figure 7.21: The effect of modifying activation flow on whether intentionally selected phonemes have significantly higher activation levels than erroneously selected phonemes, for both /k/ and /g/ productions.

models with feedback from phonemes to words. Table 7.25 reports the results of a logistic regression which examined which parameter settings make models more likely to exhibit higher pre-selection phoneme activation on intentionally selected phonemes, in comparison to erroneously selected phonemes, for both voiced and voiceless outcomes. For each specific model, comparisons of activation levels for intentional and unintentional /k/ selection were carried out separately to comparisons of activation levels for intentional and unintentional /g/ selection. In both cases, comparisons were only carried out when at least two intentional selections and two unintentional selections had taken place. To maximise the extent to which the results of the regression reflect whether a model displayed this difference, rather than whether there was enough data available for analysis, the regression only included specific models where at least two intentional /k/ selections, two unintentional /k/

Table 7.24: Binomial analysis to determine which architectures display higher pre-selection phoneme activation in intentionally selected phonemes in two-stage models with feedback from phonemes to words. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating more strongly activated intentionally selected phonemes by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	Significant activation differences		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	1423	1404	$< .001$	*
Cascading from selected Ps to Fs	2916	1418	1404	$< .001$	*
Cascading from all Ps to Fs	2916	1427	1413	$< .001$	*
Feedback from Fs to Ps	5832	2990	2972	$< .001$	*
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	2486	2284	$< .001$	*
Cascading from selected Ps to Fs	5832	2486	2273	$< .001$	*
Cascading from all Ps to Fs	5832	2471	2263	$< .001$	*
Feedback from Fs to Ps	5832	2729	2567	$< .001$	*

Key:
Ws = words, Ps = phonemes, Fs = features
Prob. = probability

selections, two intentional /g/ selections and two unintentional /g/ selections had taken place.

Figure 7.22 and table 7.25 demonstrate that models with low connection strength, a low number of steps before selection, and a high level of activation noise are most likely to display this difference. A low jolt to prime ratio also increases the probability that this difference will be significant, although the effect of this variable is weaker. There is no significant effect of decay or intrinsic noise within our results.

It would seem reasonable to suggest that high connection strengths and high numbers of steps before selection may combine to lead activation to swamp the network, so that activation levels reflect the feedback loops in the network structure more than the original activation which was input to the network. This suggestion would also help explain why feedback from phonemes to words affects whether this difference is detectable or not. As activation is added to the top of the network and the nodes involved in phoneme to word feedback are, on average, higher in the network than the nodes involved in feature to phoneme feedback, it is possible that nodes involved in phoneme to word feedback are generally more activated, which would explain why the addition of feedback from feature to phonemes does not have the

Table 7.25: Results of logistic regression model analyses using parameter values to predict which models do not show higher activation levels for intentionally selected phonemes compared to unintentionally selected phonemes, for all two-stage models with feedback from phonemes to words where at least two intentional /k/ selections, two unintentional /k/ selections, two intentional /g/ selections and two unintentional /g/ selections were recorded. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	−	31.1	1581	< .001	*
joltPrimeRatio	−	3.0	9	0.003	*
decay	−	0.2	0	0.865	
steps	−	25.0	1447	< .001	*
actiNoiseSD	+	18.6	378	< .001	*
intrinNoiseSD	−	1.0	1	0.31	

same effect. We suggest that the increased number of models found displaying this difference when high levels of activation-based noise or low jolt to prime ratios are used is mostly due to more errors occurring in these models, so that the power of these analyses is higher.

To demonstrate the effect of spreading activation parameter manipulations on trace generation, we present here results of parameter manipulation investigations for the architecture with feedback from phonemes to words without cascading from phonemes, and for the architecture with feedback from phonemes to words and from features to phonemes. This provides an overview of influences on trace generation at the featural level and at the phoneme level. Tables 7.26 to 7.27 present results of logistic regressions of the effect of parameter manipulations on the probability of trace generation; figures 7.23 and 7.25 show the effect of parameter manipulations on the probability of trace generation (where models in which no traces are generated at phoneme selection are marked separately); and figures 7.24 and 7.26 additionally show which models fail the constraints on error rate and non-contextuality. Similar patterns of results are observed for architectures with no feedback from phonemes to words, other than where noted differently.

The graphs show that in general, parameter settings which have previously been shown to lead models to generate high error rates also make models more likely to display traces which originate at feature selection, because these settings encourage featural errors to occur. These parameter settings are high decay rates, high numbers of steps before selection, high levels of intrinsic noise. Models with no cascading

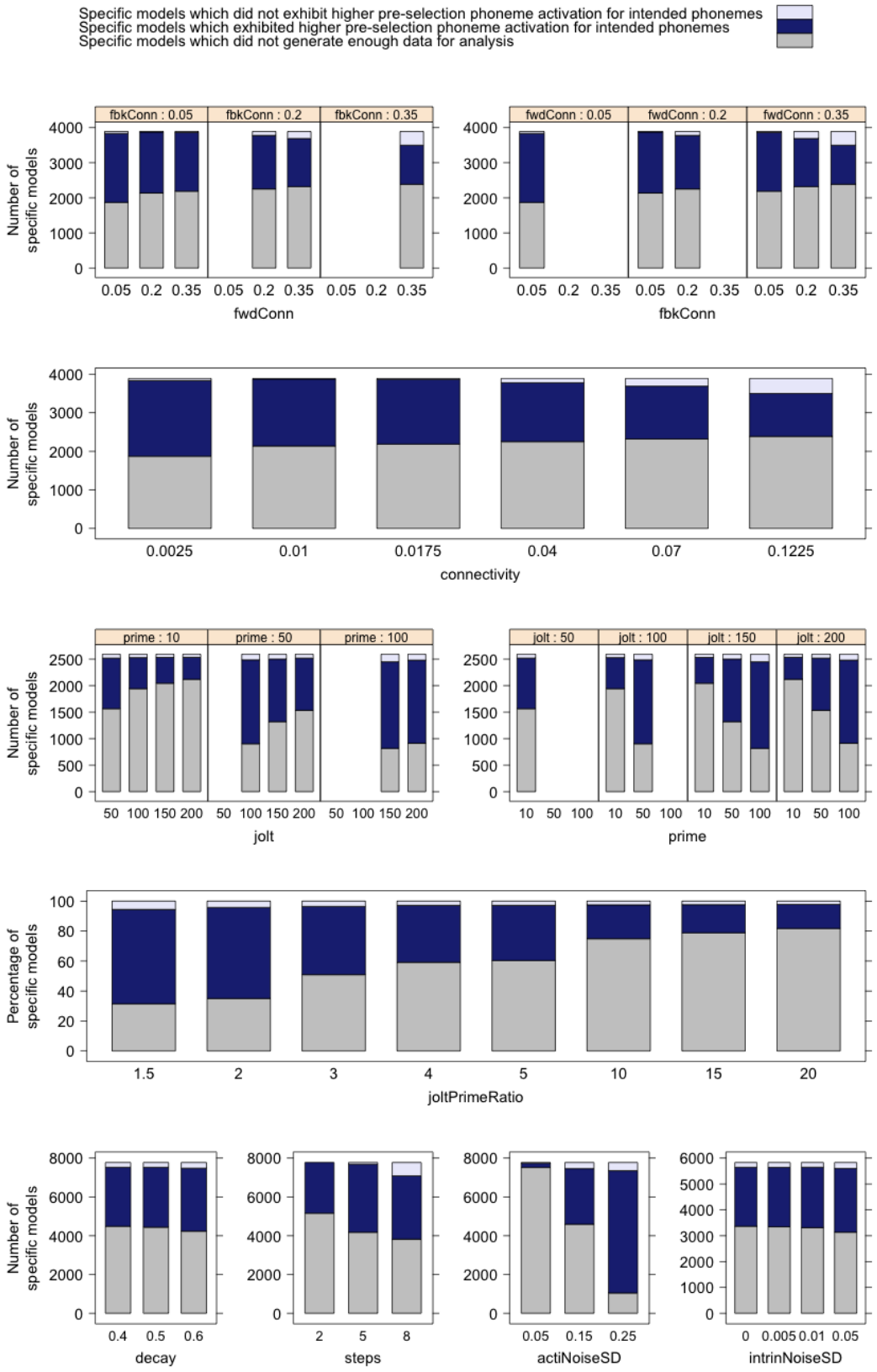


Figure 7.22: The effect of parameter manipulations on whether intentionally selected phonemes have significantly higher activation levels than erroneously selected phonemes, for both /k/ and /g/ productions, in two-stage models with feedback from phonemes to words.

from phonemes to features, or with cascading from selected phonemes to features only, in which low activation levels cause problems, additionally require low connectivity. For models with cascading from unselected phonemes to features, or with feedback from features to phonemes, where activation flooding causes problems, high connectivity is needed. High jolt to prime ratios also make models more likely to display significant traces which originate at featural selection. This is likely to be because low jolt to prime ratios particularly support errors at the phoneme level. In models with no cascading from phonemes to features, this will reduce the size of the trace effect as no trace will be present on these errors, and in other models, it will increase the chance that traces generated at the phoneme level are detected too. Where there is no cascading from phonemes, low levels of activation-based noise increase the probability of models displaying traces which originate at feature selection, because errors on highly primed competitor phonemes are further encouraged by high levels of activation-based noise. However, when there is feedback from features to phonemes, or cascading from all phonemes accompanied by feedback from phonemes to words, models with higher levels of activation-based noise are more likely to display traces which originate at feature selection, because activation cascading from a noisy phoneme-to-word feedback loop, or activation spread around the feature level due to feature-to-phoneme feedback, means that competing features are sufficiently activated for activation-based noise to substantially increase the number of feature level contextual errors which occur.

Figure 7.23 and table 7.26 correspondingly show that these are the parameter settings which make trace generation most likely in the architecture with no cascading from phonemes to features, as this architecture only generates traces on feature errors.

We suggest therefore that the main determinant of whether traces which can be shown to originate at feature selection are detected is whether enough errors are generated for these traces to be detectable, although having a sufficiently high number of contextual feature errors given the number of contextual phoneme errors may also play a role. However, as noted previously, most of these models are excluded when the constraints on error rate and non-contextuality are applied, as is particularly evident in figure 7.24, which depicts traces and model exclusion in the architecture with no cascading from phonemes to features.

Figure 7.25 shows that in the architecture with feedback from phonemes to words and from features to phonemes, when forward connection strength is high, the number of models in which significant traces are generated at feature selection

reduces as feedback connection strength increases. Previous results have shown that error rate and the proportion of non-contextual errors generated rises greatly at these parameter settings (see figures 7.4 and 7.5). This result is therefore perhaps due to a reduction in data available for analysis at these higher feedback connection strengths.

Previous sections suggested that traces stemming from phonological encoding are the main source of traces for all architectures other than the architecture with no cascading from phonemes to features. Looking across all architectures, the graphs and regressions imply that these traces are detected in models which generate high numbers of contextual phonological errors, as these provide data: i.e., specific models with low jolt to prime ratios and high activation based noise levels; but crucially, only where the parameter settings allow the subtle activation levels differences at phoneme selection to be conveyed to the featural level: i.e., specific models where the decay rate is low, there is a low number of steps before selection, and connection strength is high. On closer examination, figure 7.25 suggests that forward connection strength is more important than feedback connection strength for traces originating at the phoneme level to be detected. Further evidence for this claim is provided by the cross-tabulation of the effect of forward and feedback connection strength on the number of specific models displaying traces originating at phoneme selection shown in table 7.28. Whilst strong forward connections will aid the clear transmission of activation patterns from the phoneme level, so that the effects of small differences in phoneme activation are still discernible at feature selection, in this case feedback only serves to create more errors so that there is more data available to the analysis.

There is no significant effect of intrinsic noise on whether traces generated at phoneme selection are detected in any architecture. We also note that when jolt to prime ratio is high, much higher proportions of the specific models which generate enough data for analysis display traces (although the number of models with sufficient data decreases as jolt to prime ratio increases). As we noted when discussing the role of pre-selection phoneme activation, this is perhaps because at high jolt to prime ratios, intentionally selected phonemes receive more of a boost from the jolt than when jolt to prime ratios are low.

Figure 7.26 shows that when models which fail the constraints on error rate or non-contextuality of errors are excluded, many models in which traces generated at phoneme selection are detected with either low jolt to prime ratios or high levels of activation-based noise are excluded, because these specific models generate

Table 7.26: Results of logistic regression model analyses using parameter values to predict the occurrence of traces on both /k/ and /g/ productions, for all two-stage models with feedback from phonemes to words and no cascading from phonemes to features, where at least two intentional /k/ selections, two unintentional /k/ selections, two intentional /g/ selections and two unintentional /g/ selections were recorded. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	−	5.4	65	$< .001$	*
joltPrimeRatio	+	6.5	43	$< .001$	*
decay	+	7.2	74	$< .001$	*
steps	+	6.5	68	$< .001$	*
actiNoiseSD	−	2.2	5	0.03	*
intrinNoiseSD	+	8.2	77	$< .001$	*

too many errors. In the same way, many models with high connection strength are excluded, as we have previously shown that high connection strength leads to higher error rates and proportions of non-contextual errors, because feedback disperses activation through the network, and strong forward connections support these feedback loops. However, a high proportion of models in which traces generated at phoneme selection are detected with low decay rates and low numbers of steps before selection are not ruled out, as these parameter settings generally lead the network to generate fewer errors and lower proportions of non-contextual errors.

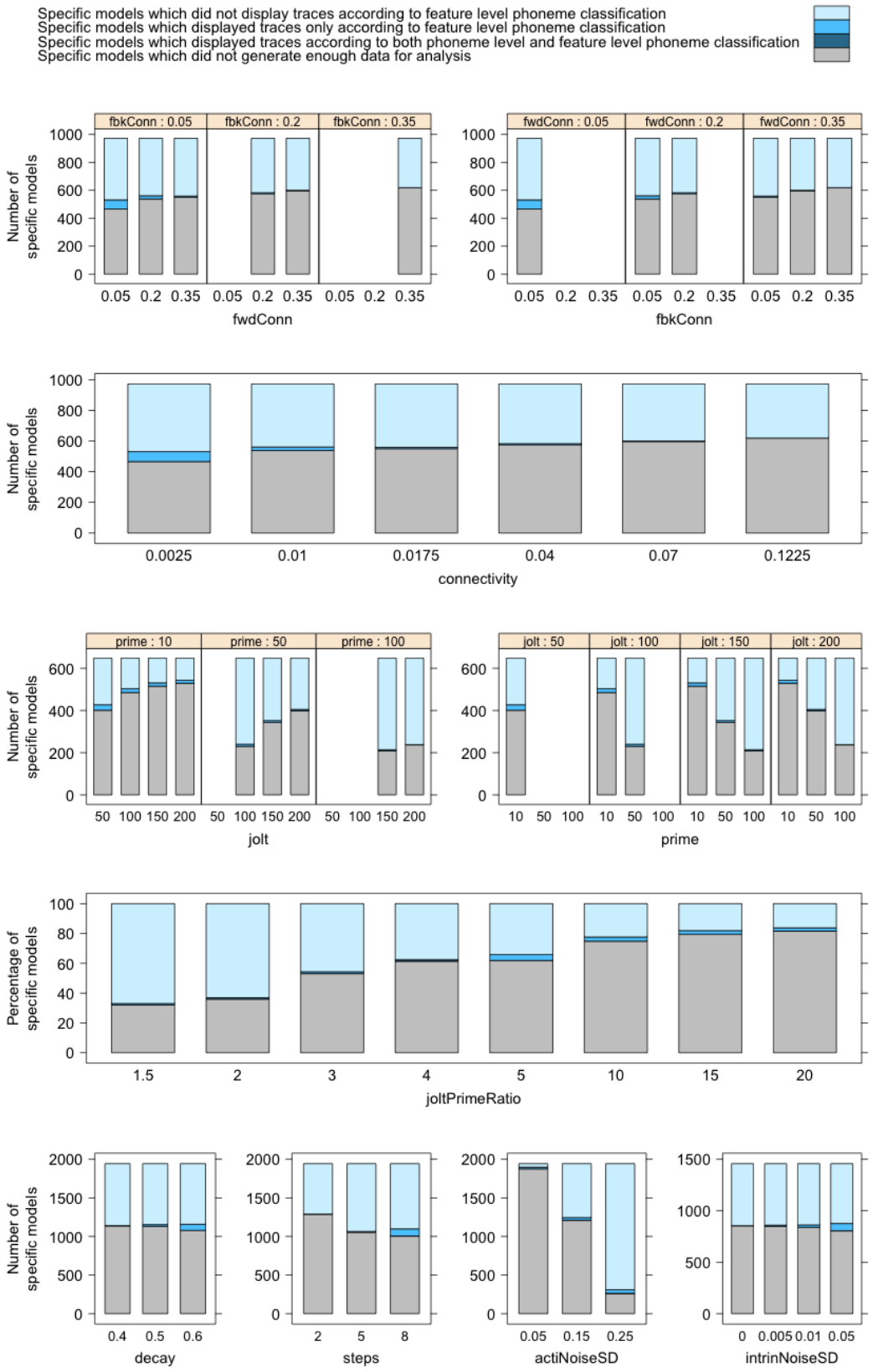


Figure 7.23: The effect of parameter manipulations on models' ability to generate traces on both /k/ and /g/ productions in two-stage models with feedback from phonemes to words and no cascading from phonemes to features, considering whether traces originated in phonological encoding or not.

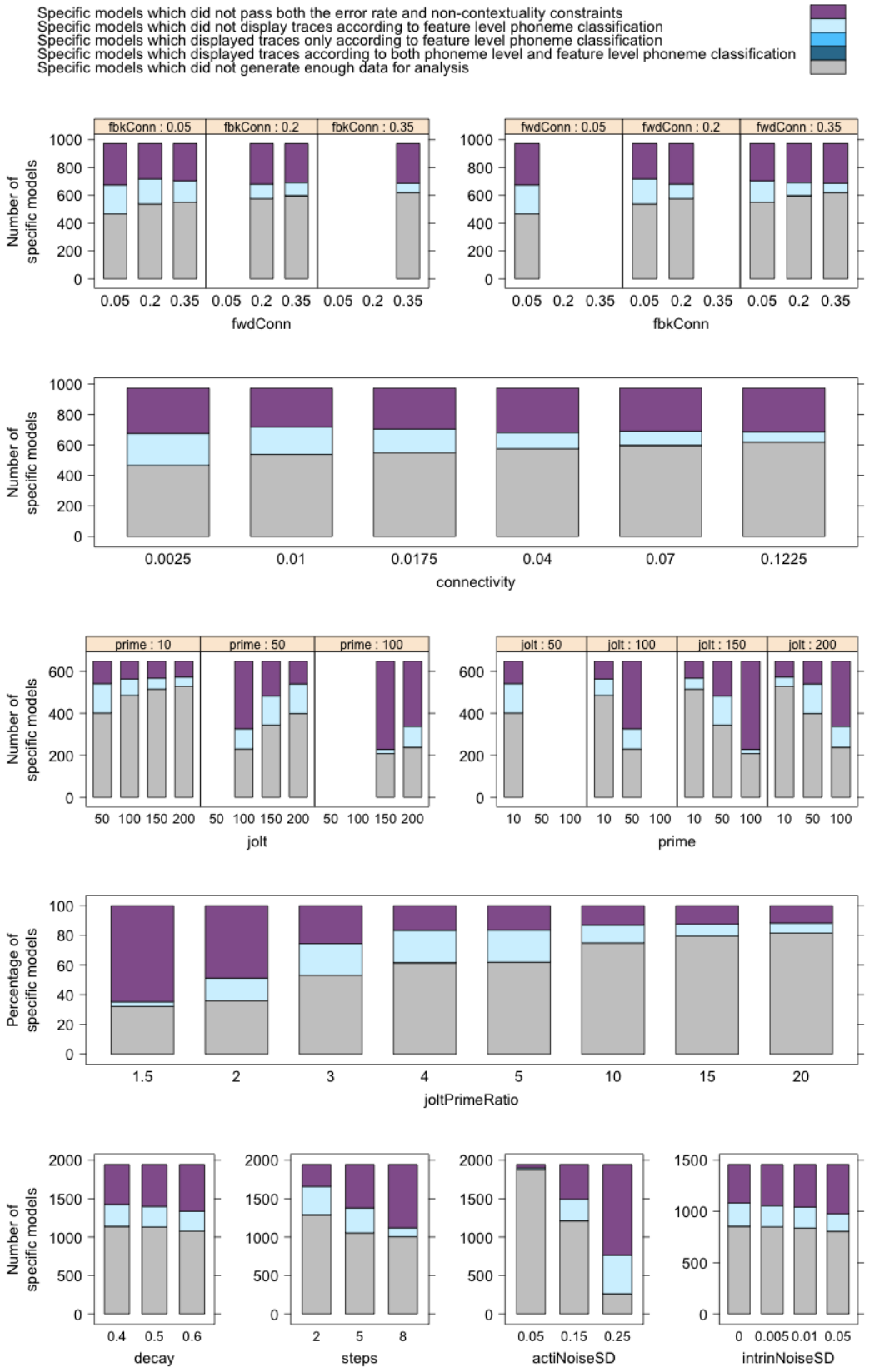


Figure 7.24: The effect of parameter manipulations on models' ability to generate traces on both /k/ and /g/ productions in two-stage models with feedback from phonemes to words and no cascading from phonemes to features, considering whether traces originated in phonological encoding or not, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

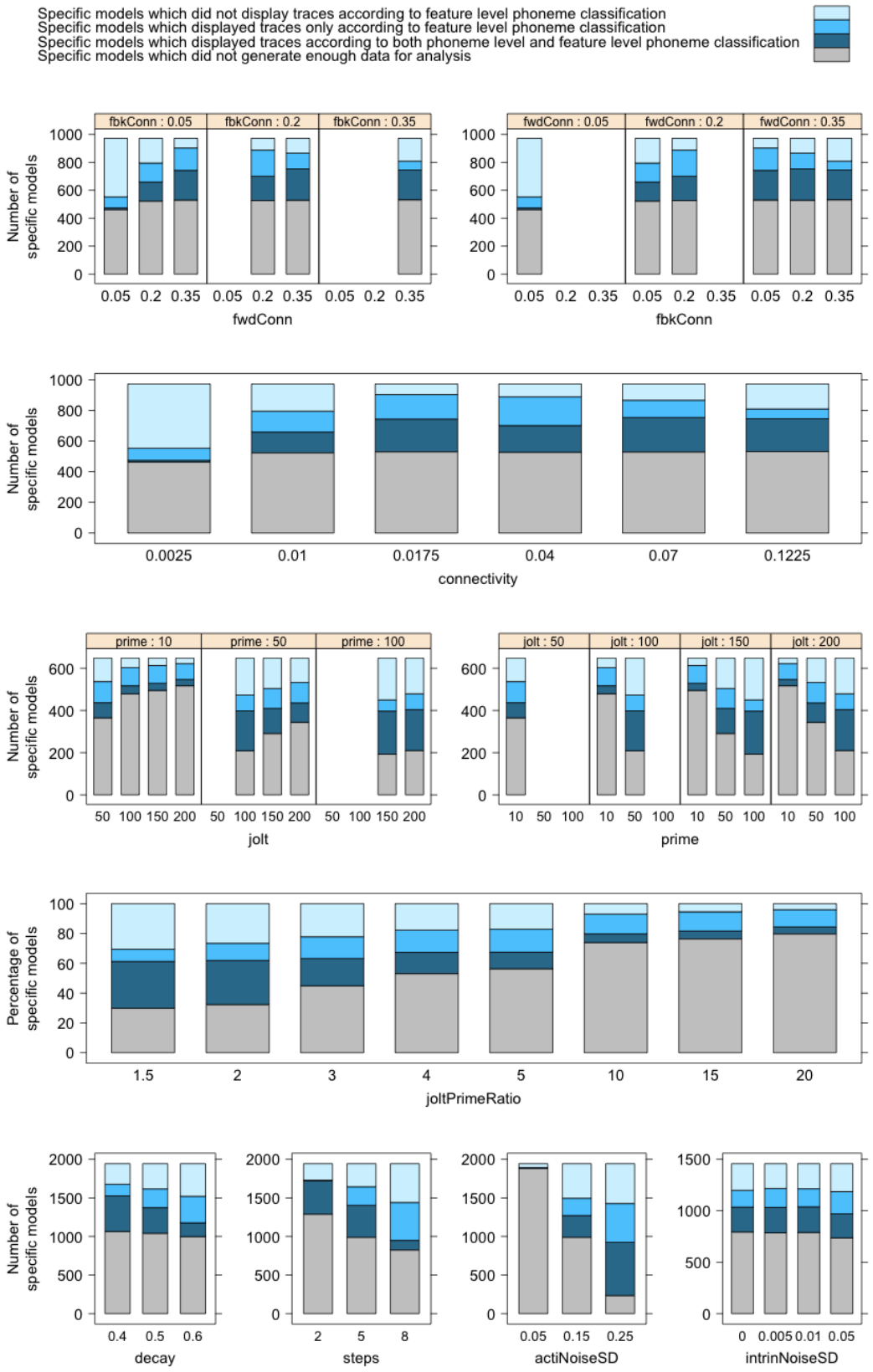


Figure 7.25: The effect of parameter manipulations on models' ability to generate traces on both /k/ and /g/ productions in two-stage models with feedback from phonemes to words and from features to phonemes, considering whether traces originated in phonological encoding or not.

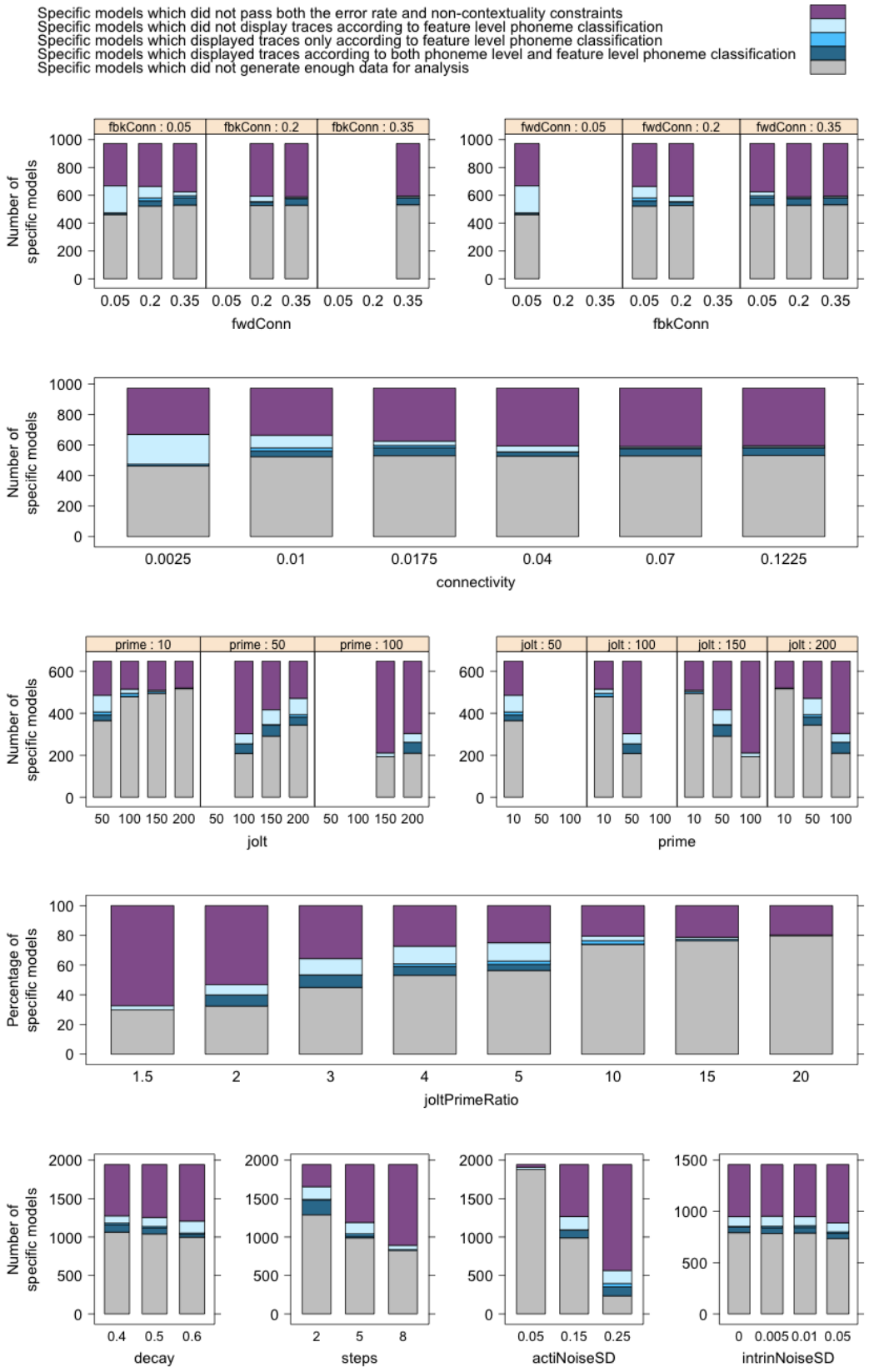


Figure 7.26: The effect of parameter manipulations on models' ability to generate traces on both /k/ and /g/ productions in two-stage models with feedback from phonemes to words and from features to phonemes, considering whether traces originated in phonological encoding or not, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

Table 7.27: Results of logistic regression model analyses using parameter values to predict the occurrence of traces on both /k/ and /g/ productions, for all two-stage models with feedback from phonemes to words and from features to phonemes, where at least two intentional /k/ selections, two unintentional /k/ selections, two intentional /g/ selections and two unintentional /g/ selections were recorded. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	**All traces**					**Phoneme selection traces**				
	Direction	Z	LRT	P (χ^2)		Direction	Z	LRT	P (χ^2)	
connectivity	+	7.2	54	< .001	*	+	12.5	168	< .001	*
joltPrimeRatio	+	5.7	35	< .001	*	−	4.8	24	< .001	*
decay	−	6.1	37	< .001	*	−	15.7	286	< .001	*
steps	−	6.0	36	< .001	*	−	21.0	568	< .001	*
actiNoiseSD	+	7.3	55	< .001	*	+	6.6	46	< .001	*
intrinNoiseSD	+	0.7	0	0.507		−	1.2	2	0.215	

Table 7.28: Cross-tabulation of the effect of forward and feedback connection strength on the number of specific models displaying traces originating at phoneme selection, for all models with feedback from phonemes to words, and either cascading from selected phonemes only, or cascading from all phonemes, or feedback from features to phonemes.

	fbkConn		
	0.05	0.2	0.35
fwdConn			
0.05	27	-	-
0.2	306	344	-
0.35	486	466	447

Transcribed lexical bias and phonological similarity effects, and traces of intended phonemes on errors

Finally, we combined the data from this simulation with data reported in section 7.3 to determine whether any of the two-stage architectures could simultaneously account for the transcribed lexical bias effect, the transcribed phonological similarity effect, and traces of intended phonemes on both errorful voiced productions and errorful voiceless productions. Figure 7.27 shows that a number of models with feedback from phonemes to words and either cascading from all phonemes or feedback from features to phonemes do indeed show all four effects, and the binomial analysis reported in table 7.29 confirms that for both of these architectures, there are more models reporting all four effects than would be predicted by chance. This

is particularly notable, as it shows that this analysis methodology loses power when by used to account for too many simultaneous results. The per specific model probability that at least one significant result was a Type I error is now $1-0.95^4 = 0.185$ (to 3 d.p.). The evidence for the model's ability to capture all four effects is both numerically and statistically weaker once models which fail the constraints on error rate and non-contextuality are excluded however. For the architecture with feedback from phonemes to words and cascading from all phonemes to features, only 39 models exhibit all four effects, of the 685 which generate errors for analysis and pass the constraints; and for the architecture with feedback from phonemes to words and from features to phonemes, only 24 models exhibit all four effects, of the 634 which generate errors for analysis and pass the constraints, as shown in table 7.30.

We examined the effect of the spreading activation parameters on the network's behaviour to see if this could provide further insight into why the error rate and non-contextuality constraints caused a problem. Analyses showed that the parameters had a similar effect on the architecture with feedback from phonemes to words and cascading from all phonemes to features, so we collapse over the two architectures for this analysis. Figure 7.29 and the logistic regression reported in table 7.31 show that specific models with high connection strengths, high numbers of steps before selection, high levels of activation-based noise, low decay rate and a high jolt to prime ratio are particularly successful at generating all four effects. However, as shown in section 7.2, models with high connection strength, a high number of steps before selection and high levels of activation-based noise lead to high error levels (and high proportions of non-contextual errors). This both helps explain why these parameter settings make it more likely that both effects are detected (more error data to increase the power of the analysis), but also means that these models are particularly likely to be ruled out by the constraints on error rate and non-contextuality, as reflected in figure 7.30. Given the conflict between the need to ensure that model error rates are reasonable in comparison to human error rates, and the need for errors for analysis where a higher number of errors will increase the power of an analysis, it would in future be worth running these simulations with more trials, so that specific models with a lower error rate have a better chance to generate enough data.

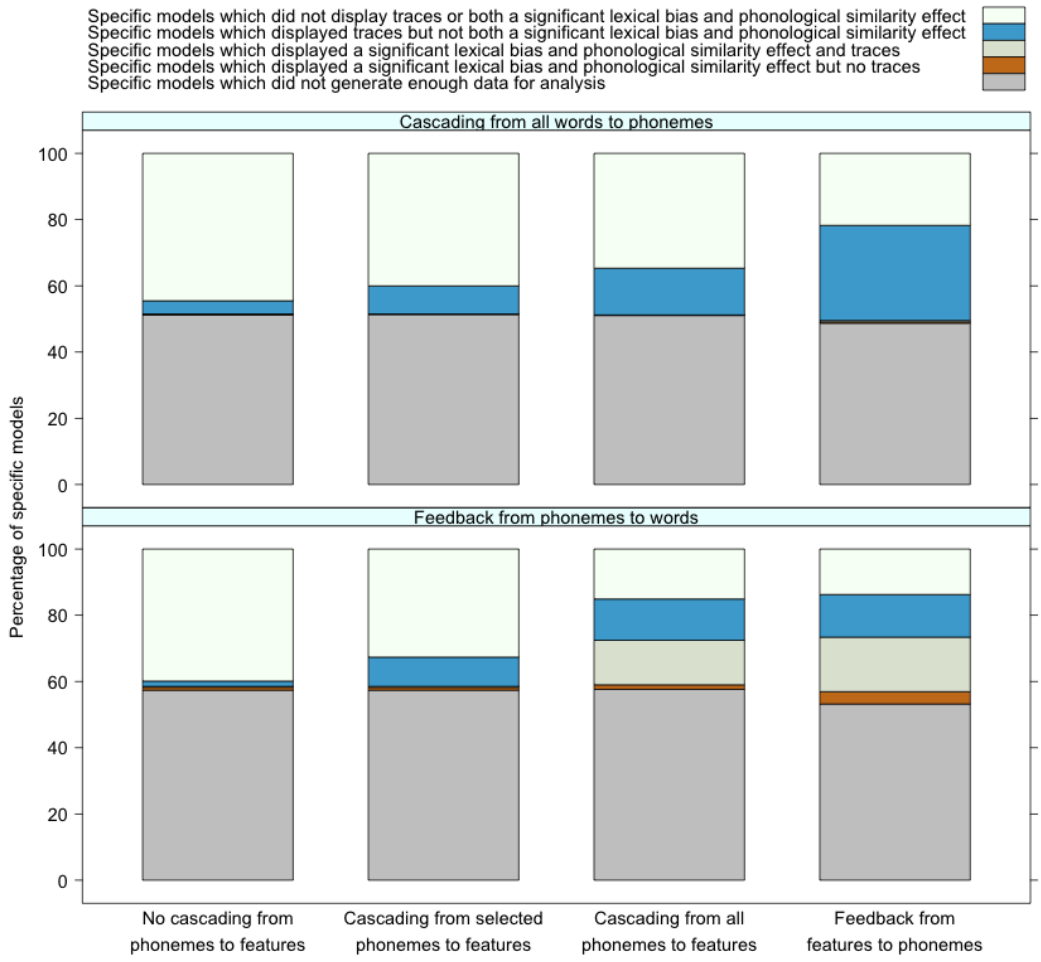


Figure 7.27: The effect of modifying activation flow on two-stage models' ability to simultaneously display the lexical bias effect, the phonological similarity effect and generate traces on both /k/ and /g/ productions

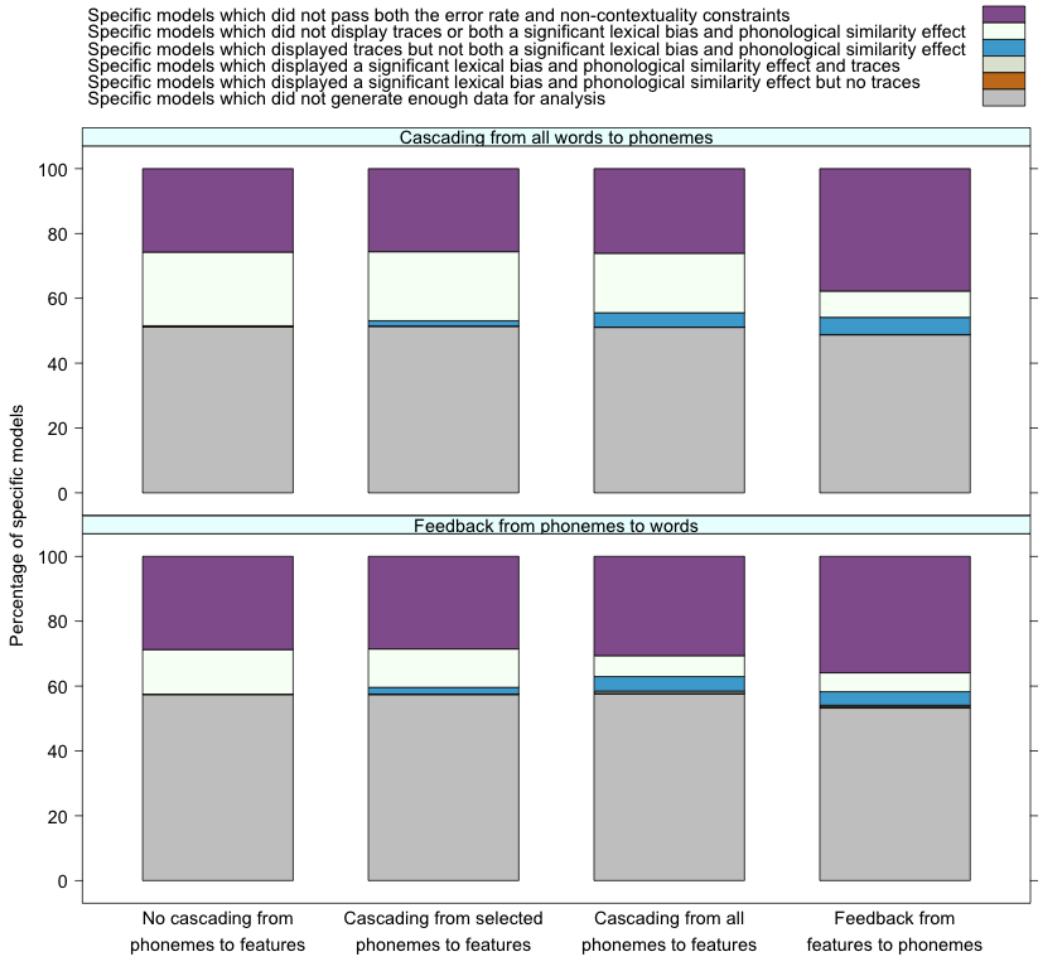


Figure 7.28: The effect of modifying activation flow on two-stage models' ability to simultaneously display the lexical bias effect, the phonological similarity effect and generate traces on both /k/ and /g/ productions, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

Table 7.29: Binomial analysis to determine which two-stage architectures simultaneously exhibit traces of intended voiced phonemes on voiceless productions, traces of intended voiceless phonemes on voiced productions, lexical bias effects and the phonological similarity effect. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models simultaneously generating these effects by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	Significant LB, PS and traces		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	1423	2	$> .9$	
Cascading from selected Ps to Fs	2916	1418	2	$> .9$	
Cascading from all Ps to Fs	2916	1427	2	$> .9$	
Feedback from Fs to Ps	5832	2990	24	$> .9$	
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	2486	14	$> .9$	
Cascading from selected Ps to Fs	5832	2486	16	$> .9$	
Cascading from all Ps to Fs	5832	2471	784	$< .001$	*
Feedback from Fs to Ps	5832	2729	960	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

LB = lexical bias effect, PS = phonological similarity effect

Prob. = probability

Table 7.30: Binomial analysis to determine which two-stage architectures simultaneously exhibit traces of intended voiced phonemes on voiceless productions, traces of intended voiceless phonemes on voiced productions, lexical bias effects and the phonological similarity effect, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models simultaneously generating these effects by chance.

	Specific model counts				**Prob.**
	Total	Excluded	Sufficient data	Significant LB, PS and traces	
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	751	672	0	$> .9$
Cascading from selected Ps to Fs	2916	747	671	0	$> .9$
Cascading from all Ps to Fs	2916	761	666	0	$> .9$
Feedback from Fs to Ps	5832	2203	787	0	$> .9$
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	1676	810	0	$> .9$
Cascading from selected Ps to Fs	5832	1665	821	3	$> .9$
Cascading from all Ps to Fs	5832	1786	685	39	$> .9$
Feedback from Fs to Ps	5832	2095	634	24	$> .9$

Key:

Ws = words, Ps = phonemes, Fs = features

LB = lexical bias effect, PS = phonological similarity effect

Prob. = probability

Table 7.31: Results of logistic regression model analyses using parameter values to predict the simultaneous occurrence of the lexical bias effect, the phonological similarity effect and traces on both /k/ and /g/ productions, for all two-stage models with feedback from phonemes to words and either cascading from all phonemes or feedback from features to phonemes where in the trace experiment, at least two intentional /k/ selections, two unintentional /k/ selections, two intentional /g/ selections and two unintentional /g/ selections were recorded. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	+	13.6	189	$< .001$	*
joltPrimeRatio	+	7.0	50	$< .001$	*
decay	−	2.6	7	0.009	*
steps	+	22.4	587	$< .001$	*
actiNoiseSD	+	10.2	111	$< .001$	*
intrinNoiseSD	+	0.5	0	0.585	

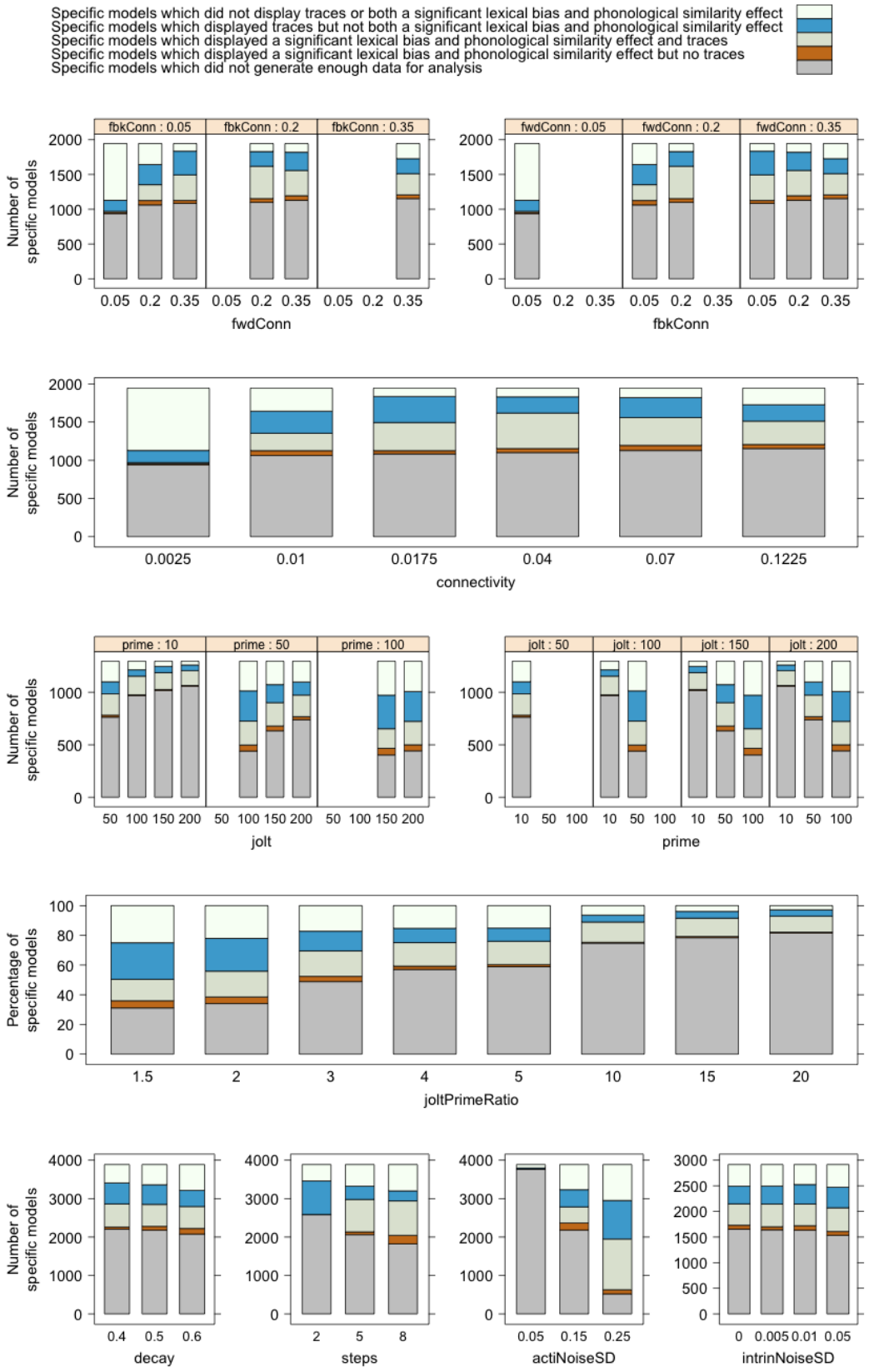


Figure 7.29: The effect of parameter manipulations on models' ability to simultaneously display the lexical bias effect, the phonological similarity effect and generate traces on both /k/ and /g/ productions, for two-stage models with feedback from phonemes to words and either cascading from all phonemes or feedback from features to phonemes.

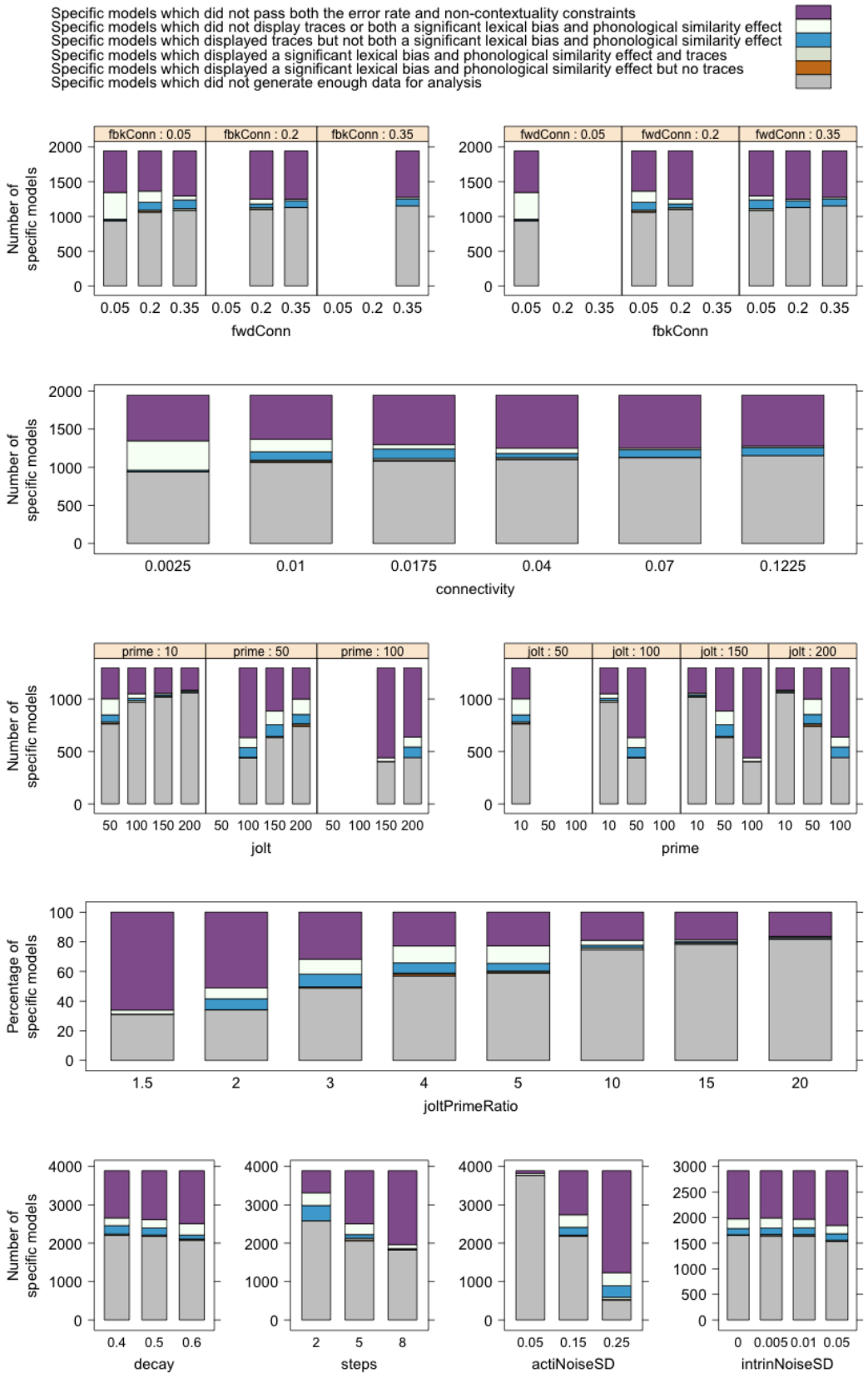


Figure 7.30: The effect of parameter manipulations on models' ability to simultaneously display the lexical bias effect, the phonological similarity effect and generate traces on both /k/ and /g/ productions, for two-stage models with feedback from phonemes to words and either cascading from all phonemes or feedback from features to phonemes, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

7.4.3 Conclusions

Goldrick and Blumstein (2006) argued that traces of intended phonemes on errorful productions reflect activation cascading from intended but unselected phonemes, and that a model of word production would therefore require cascading from all phonemes to explain this evidence. In contrast to Goldrick and Blumstein's (2006) predictions, we suggested that all architectures should be able to generate traces, as there are actually another two mechanisms by which traces may be generated: traces of the intended phoneme due to errors at the featural level, a mechanism which we predicted would work for all architectures; and traces of the intended phoneme due to weaker activation of unintended but selected phonemes, a mechanism which we argued would operate in all architectures with cascading from selected phonemes, without a requirement for cascading from unselected phonemes.

When analysing voiced outcome productions (i.e., /g/ → [g] compared to /k/ → [g]) and voiceless outcome productions (i.e., /k/ → [k] compared to /g/ → [k]) separately, we found that models from all of the architectures were able to display traces, confirming our prediction. Furthermore, the number of models displaying significant traces increased as the interactivity of the activation flow between phonemes and features increased, because more interactive models had more mechanisms for generating traces available to them. The model with feedback from features to phonemes could only generate traces in the same ways that the model with cascading from all phonemes to features could, but in this model, more errors were generated such that there was more data available to the statistical tests for the presence of traces. In addition, there were more models with significant traces on voiceless productions when phoneme-to-word feedback was included in the model, as there were more voiceless outcome errors.

However, exclusion of all specific models which failed the constraints on error rate and non-contextuality led most specific models in which traces originated from errors at feature selection only to be ruled out. This was a particular problem for the architecture with no cascading from phonemes, in which no other mechanism of trace generation can operate. Similarly, when we examined which specific models show significant traces on both voiced productions and voiceless productions, we found that too few models with no cascading from phonemes to features generated traces on both types of production to rule out a null hypothesis of chance trace generation. We argued that these results echoed the results of section 7.3, in which it was shown that where there is no influence of the prime activation on featural

errors (a situation which is frequently due to restricted activation flow), feature errors are essentially random, such that a high proportion of non-contextual errors are generated. This in turn means that for enough contextual errors to be generated for effects on these errors to be detected, the overall error rate must be very high. Very few models generate high enough error rates for traces on contextual errors to be detected, which is a problem when the lower power multiple effects binomial analysis is used; and models which do generate enough errors are ruled out by the constraints on error rate and non-contextuality of errors. Future work will verify that these problems are due to priming being applied at the the word level, and do not arise if priming is applied at the featural level, to simulate perseverative influences in tongue twisters for example.

There is no architecture for which all specific models with enough data for analysis generate traces however. We verified that in nearly all models, intentionally selected phonemes were more activated than erroneously selected phonemes. A few models with phoneme-to-word feedback, high connection strength, a high number of steps before selection and a high level of activation based noise did not show this difference, which we suggested was due to activation flooding the network such that the identity of the phoneme which originally received the jolt activation became irrelevant. However, because so few models did not show this difference, this did not clearly explain the large number of models not displaying traces.

We therefore directly investigated the effect of spreading activation parameter manipulations on trace generation. It was shown that traces which originate at featural errors are detected in specific models with high decay rates, a high number of steps before selection, high intrinsic noise, and either low connectivity or high connectivity, depending on the model architecture. These are models in which error rates are high, so featural errors are likely. A high jolt to prime ratio also makes it more likely that a model is classified as generating traces at the featural level only, as low jolt to prime ratios encourage phoneme level errors, which will either reduce the size of the effect as not trace will be present, or will lead to the model being classified as generating traces at the phoneme level too. Traces originating at phonological encoding were detected on models in which a high number of errors are generated at the phoneme level, as these provide data: i.e., specific models with low jolt-prime ratios and high activation based noise levels; but only where the parameter settings permit the subtle activation levels differences at phonological encoding to be conveyed to the featural level: i.e., specific models with a low decay rate, a low number of steps before selection, and a high forward connection strength. The fact that

different parameter settings were required for different accounts of trace generation gives weight to our original concerns that investigating the behaviour of different architectures at arbitrarily chosen parameter settings would not have constituted a fair test.

We finally examined whether any two-stage architecture can account for the transcribed lexical bias effect, the transcribed phonological similarity effect and traces of intended phonemes on unintended phonemes for voiced and voiceless productions. We found that models with phoneme-to-word feedback and either cascading from all phonemes or feedback from features to phonemes were capable of demonstrating all four effects. Testing for four effects means that lack of power due to compensation for Type I inflation is quite severe, so it is notable that these architectures still pass this test. Once models which fail the constraints on error rate or non-contextuality are excluded however, an extremely low number of models demonstrate all four effects, and no statistical evidence is found to confirm that any architecture can account for all effects simultaneously. We propose that a large part of this problem is due to a very high error rate being required for all four effects to be detected without any Type II errors occurring. As error analyses require errors to be generated, and higher numbers of errors result in more powerful analyses, it would in the future be worth running further simulations with more trials, so that specific models with lower error rates (and which consequently pass the constraint on error rate) have a chance to generate more data.

7.5 Goldrick and Blumstein's (2006) acoustic evidence of a lexical bias effect on traces

The previous sections have demonstrated that both the phonological similarity effect in transcribed speech error evidence and Goldrick and Blumstein's (2006) VOT evidence of traces of intended phonemes on erroneous production can be accounted for by a model with no cascading from phonemes. In this final simulation of the chapter, we investigate whether Goldrick and Blumstein's (2006) finding of a lexical bias on VOT traces places any stronger constraints on the nature of activation flow between phonemes and features. Goldrick and Blumstein (2006) again claimed that this result demonstrated cascading from all phonemes, reflecting suppression of the activation cascading from the intended phoneme in the lexical error outcome condition. However, we argued that extra activation cascading from the unintentionally selected phoneme in the lexical error outcome condition is sufficient to explain this

result, such that any architecture with cascading from selected phonemes would be able to account for Goldrick and Blumstein's (2006) post-hoc finding. Common to both Goldrick and Blumstein's (2006) argument and our own hypothesis is the assumption that models with no cascading from phonemes will not be able to account for this effect, as in this architectures there is no way for the lexicality of the error outcome to affect errors at the featural level. We seek to confirm these hypotheses with the simulations reported here.

7.5.1 Simulation methodology

The results reported here come from the simulations reported in section 7.3, in which the lexicality of the error outcome and phonological similarity of the target onset and competing onset were manipulated. Different analyses of the output of these simulations allow us to determine which specific models simulate the lexical bias on traces reported by Goldrick and Blumstein (2006).

Model configuration

As reported in section 7.3, we examined the behaviour of all 37,908 two-stage models.

Model task and lexicon

To recap, the model's task was to produce single words, while competitor words were primed. Materials were designed in which lexicality of the error outcome was manipulated, as well as phonological similarity of the target and competing onset, although the phonological similarity manipulation was not relevant here. These materials are described in full in section 6.2.2. In this study, we consider behaviour from the simulation using the material set in which voicing of the target and competitor onset always differs, to allow us to simulate Goldrick and Blumstein's (2006) VOT experiment. The 100 word lexicon used is also described in section 6.2.2.

Model output interpretation

To compare model behaviour to Goldrick and Blumstein's (2006) finding of a lexical bias on VOT traces, output was first categorised as a correct production, a contextual error, or a non-contextual error depending on the features selected at the end of subphonemic processing. The VOT of correct productions and contextual errors was then calculated in the manner outlined in chapter 3.

To verify that traces were also produced with this different material set and lexicon, a t-test comparing the VOT of all intended voiced productions to all unintended voiced productions was carried out, and a similar t-test was executed for voiceless productions.

To determine whether a lexical bias was present on traces, we again considered voiced and voiceless productions separately. Here we take voiced outcome productions as an example. For the lexical error outcome and the non-lexical error outcome conditions separately, we calculated the average VOT for an intended voiced production. The difference between the VOT of each unintended voiced production and the average VOT for an intended voiced production was then calculated. We refer to this difference here as a trace. Lastly, a t-test comparison of the traces in the lexical condition and the traces in the non-lexical condition was carried out to determine whether traces in the lexical condition were smaller than traces in the non-lexical condition as in Goldrick and Blumstein's (2006) results.

All t-tests were only carried out if there were at least two contextual errors and two correct productions in each of the lexicality conditions.

7.5.2 Simulation results

We first verified that the results reported in section 7.4 were replicated with the different materials and lexicon used in the current simulation. We then investigated which architectures allowed a lexical bias on traces to be displayed. Finally, we examined the effect of spreading activation parameter manipulations on the generation of lexical bias effects on VOT traces. We report these results in this section.

Replication of the original VOT trace simulation results

Investigations of which architectures demonstrated traces of intended phonemes on unintended productions gave almost identical results to those found in section 7.4. A binomial analysis (with Bonferroni corrected $\alpha = 0.00625$) showed that when considering productions on voiced and voiceless onsets separately, all architectures could generate traces regardless of word-to-phoneme or phoneme-to-feature activation flow (all ps < 0.001). However, when the constraints on error rate and non-contextuality of errors are applied, no statistical evidence was found to show that the architecture with no feedback from phonemes to words and no cascading from phonemes can account for traces on either voiced outcome productions

(architecture with no feedback from phonemes to words and no cascading from phonemes: $p = 0.034$; all other architectures: $p < 0.001$) or voiceless outcome productions (architecture with no feedback from phonemes to words and no cascading from phonemes: $p = 0.022$; all other architectures: $p < 0.003$). In section 7.4, evidence was found that this architecture can generate traces on voiceless outcome productions when the constraints are applied, but no evidence was found for effect generation in the architecture with feedback from phonemes to words and no cascading from phonemes. These results therefore serve to support our argument that feedback from phonemes to words is not exerting a great effect on trace generation in the architecture with no cascading from phonemes to words. Rather, in the current implementation where prime activation is applied at the word level, priming cannot reach the featural level to support contextual error generation. A low number of contextual errors generated reduces the power of the trace analysis. Specific models which generate enough errors for trace effects to be detected will also generate very large numbers of non-contextual errors and are consequently likely to be ruled out by the constraints on error rate and non-contextuality. Type II errors are therefore more likely on any architecture with no cascading from phonemes, regardless of activation flow between words and phonemes.

As in section 7.4, paucity of data for trace effect detection in architectures with no cascading from phonemes in the current implementation also means that no statistical evidence is found for the architecture's ability to generate trace effects when both voiced and voiceless outcome productions are considered (both no cascading from phonemes architectures: $p > 0.9$; all other architectures: $p < 0.001$). This is partially because the power of the binomial analysis is weakened in this multiple effect detection scenario. In addition, when models failing the constraints on error rate and non-contextuality of errors are excluded, no statistical evidence is found that architectures with no feedback from phonemes to words and cascading from selected phonemes only can account for the trace effect on both voiced and voiceless outcome productions (architecture with no feedback from phonemes to words and cascading from selected phonemes only: $p > 0.8$; both no cascading from phonemes architectures: $p > 0.9$; all other architectures: $p < 0.001$). As in section 7.4, we argue that this is because the phoneme selection error rate is low in this architecture, reducing the number of models for which traces due to difference in activation levels of intended and unintended phonemes can be detected.

Finally, we note that as in our previous results, when feedback from phonemes to words is present, there are numerically more voiceless outcome traces than voiced

outcome traces. This is because the voiceless phonemes occur as onsets, on average, more frequently than the voiced onsets in the model lexicon. This increases the probability of voiceless outcome errors, thereby also increasing the power of the tests for the presence of traces for voiceless outcomes. Similarly, there are more voiced outcome traces than voiceless outcome traces when feedback from features to phonemes is present, as there are more onsets which are voiced than voiceless onsets, increasing the probability of erroneous selection of the voiced feature.

In conclusion, using different materials and lexicon, we replicate the results we reported in section 7.4 from a simulation of Goldrick and Blumstein's (2006) VOT trace evidence, demonstrating that models with no cascading from phonemes can account for these results. However, a lack of priming support for contextual error generation at the feature level in the current implementation and the resulting high proportion of non-contextual errors generated at the featural level causes some problems which should be addressed in future simulations applying priming at the featural level.

Lexical bias on VOT traces

We then investigated which specific models demonstrate a lexical bias on VOT traces, as reported by Goldrick and Blumstein (2006), such that traces of intended phonemes on erroneous productions are smaller in the lexical outcome condition than in the non-lexical condition. As this result was from a post-hoc analysis of Goldrick and Blumstein's (2006) evidence for traces of intended phonemes on erroneous productions in humans, we only considered the behaviour of specific models which demonstrate such traces. Our results showed that, contrary to both Goldrick and Blumstein's (2006) and our own predictions, all architectures with feedback from phonemes to words could account for lexical bias on VOT traces, as depicted in figure 7.31 and confirmed by the analyses reported in tables 7.32 and 7.33.

In light of this finding, it becomes clear that we had overlooked a mechanism by which the architecture with no cascading from phonemes could account for this effect. Errors can occur at both the phoneme level and the feature level, but in this architecture, only errors at the feature level will show a trace of the intended phoneme. However, more errors will occur at the phoneme level when the error outcome is lexical. Errors at the feature level are not affected by this variable. This means that a higher proportion of errors in the lexical outcome condition are generated at the phoneme level, in comparison to the non-lexical outcome condition.

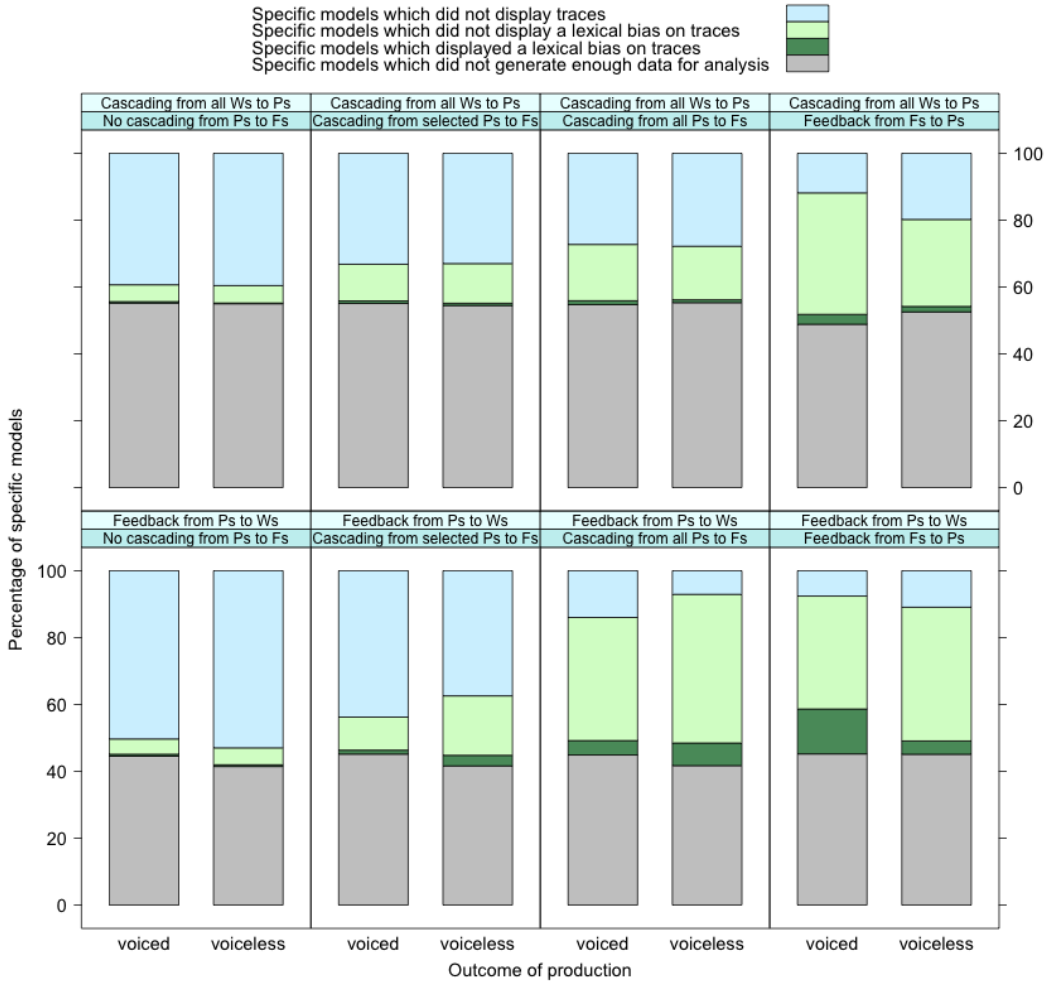


Figure 7.31: The effect of modifying activation flow on whether two-stage models generate smaller traces on lexical than for non-lexical error outcome productions of voiceless and voiced onset consonants.
Key: Ws = words, Ps = phonemes, Fs = features

Table 7.32: Binomial analysis to determine which two-stage architectures generate smaller traces of intended voiceless phonemes on voiced productions for lexical than for non-lexical error outcome productions. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating lexical bias effects by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data and traces	Lexical bias on traces		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	160	13	0.03	
Cascading from selected Ps to Fs	2916	342	22	0.094	
Cascading from all Ps to Fs	2916	525	34	0.054	
Feedback from Fs to Ps	5832	2292	173	$< .001$	*
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	301	35	$< .001$	*
Cascading from selected Ps to Fs	5832	645	69	$< .001$	*
Cascading from all Ps to Fs	5832	2399	252	$< .001$	*
Feedback from Fs to Ps	5832	2753	784	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

Table 7.33: Binomial analysis to determine which two-stage architectures generate smaller traces of intended voiced phonemes on voiceless productions for lexical error outcome productions than non-lexical error outcome productions. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating lexical bias effects by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data and traces	Lexical bias on traces		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	158	6	0.68	
Cascading from selected Ps to Fs	2916	366	20	0.29	
Cascading from all Ps to Fs	2916	492	25	0.414	
Feedback from Fs to Ps	5832	1614	96	0.038	
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	325	30	0.001	*
Cascading from selected Ps to Fs	5832	1222	185	$< .001$	*
Cascading from all Ps to Fs	5832	2988	395	$< .001$	*
Feedback from Fs to Ps	5832	2565	232	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

As phoneme level errors show no trace of the intended phoneme, the overall trace size in the lexical level will be smaller.

This explanation clearly predicts a bimodal distribution of traces in the lexical condition, and it is not clear from Goldrick and Blumstein's (2006) report whether such a distribution is found. Nevertheless, this result demonstrates that the simple observation that traces are on average smaller in the lexical outcome condition is not enough to distinguish between these models. Furthermore, it highlights an oversight in our own pen and paper reasoning about the behaviour of the model and provides confirmation of the need for explicit modelling.

Our results also suggested that the architecture with no feedback from phonemes to words and feedback from features to phonemes displays a lexical bias on traces, but only when the outcome of the error is voiced, as can be seen in tables 7.32 and 7.33. In contrast to our other unexpected result, we found no obvious explanation for why any architecture with no feedback from phonemes to words would generate a lexical bias on traces. No lexical bias on categorised errors was found for this architecture in section 7.3, and indeed no significant result was found for this architecture for a lexical bias on voiceless outcome traces either. We conclude that this finding is therefore most likely to either represent a Type I error, or be the result of a confound in the materials that we have not located (where the design of these materials is described in section 6.2.2). Replication of this experiment would help distinguish between these possibilities.

Figure 7.31 shows however that overall, a very low number of specific models demonstrate a lexical bias on traces. Given that a lexical bias on traces is a modification of another effect, it is quite possible that the final effect is weak. In this case, there is possibly a greater chance that the effect will be detected on specific models which generate many errors, thereby boosting the power of the test for the effect's presence. It is therefore not surprising that, as can be seen in figure 7.32, very few specific models which demonstrate this effect remain when specific models which fail the constraints on error rate or non-contextuality of errors are ruled out. As a result, table 7.34 shows that there is no statistical evidence that models of any architectures are able to display a lexical bias effect on traces on voiced error outcomes without failing these constraints, and table 7.35 demonstrates that for voiceless outcome errors, there is only significant evidence for architectures with feedback from phonemes to words and either cascading from selected phonemes only, or cascading from all phonemes.

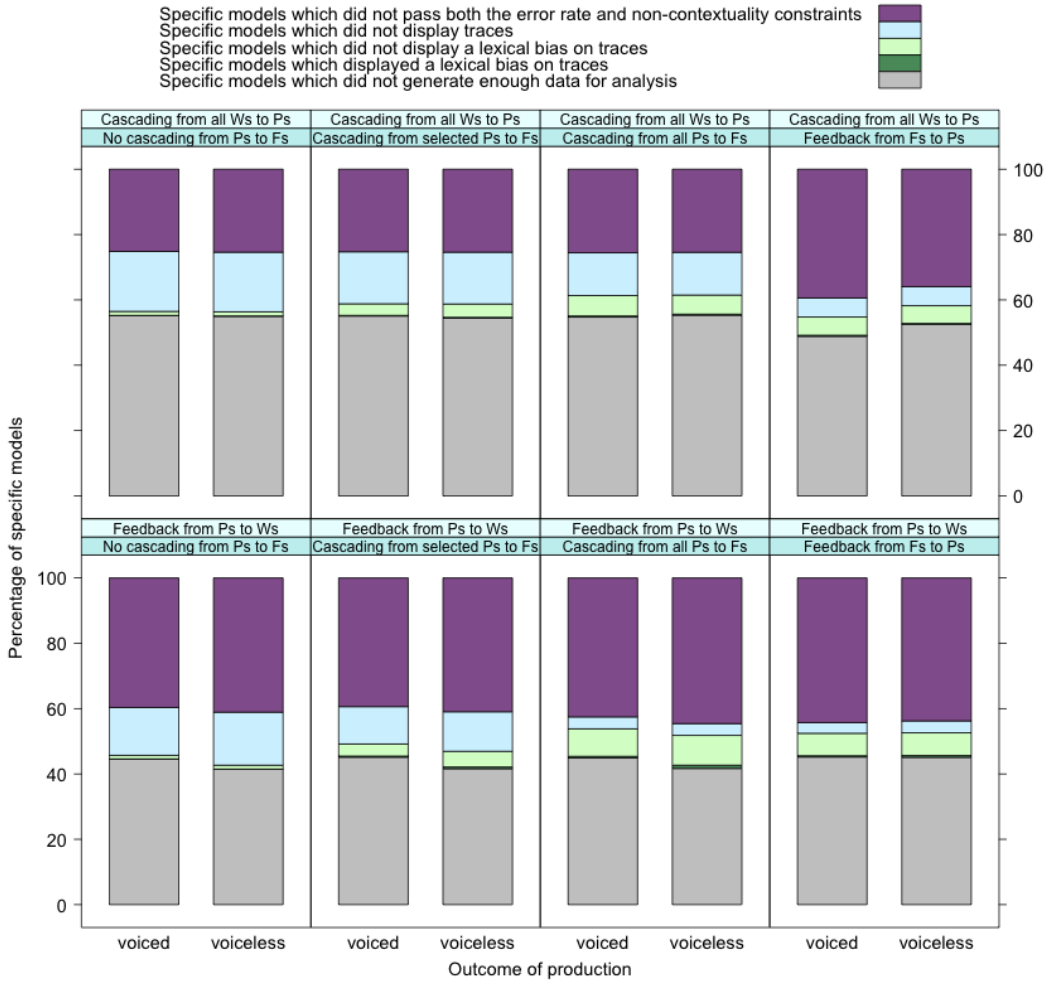


Figure 7.32: The effect of modifying activation flow on whether two-stage models generate smaller traces on lexical than for non-lexical error outcome productions, on voiceless and voiced productions in two-stage models, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.
Key: Ws = words, Ps = phonemes, Fs = features

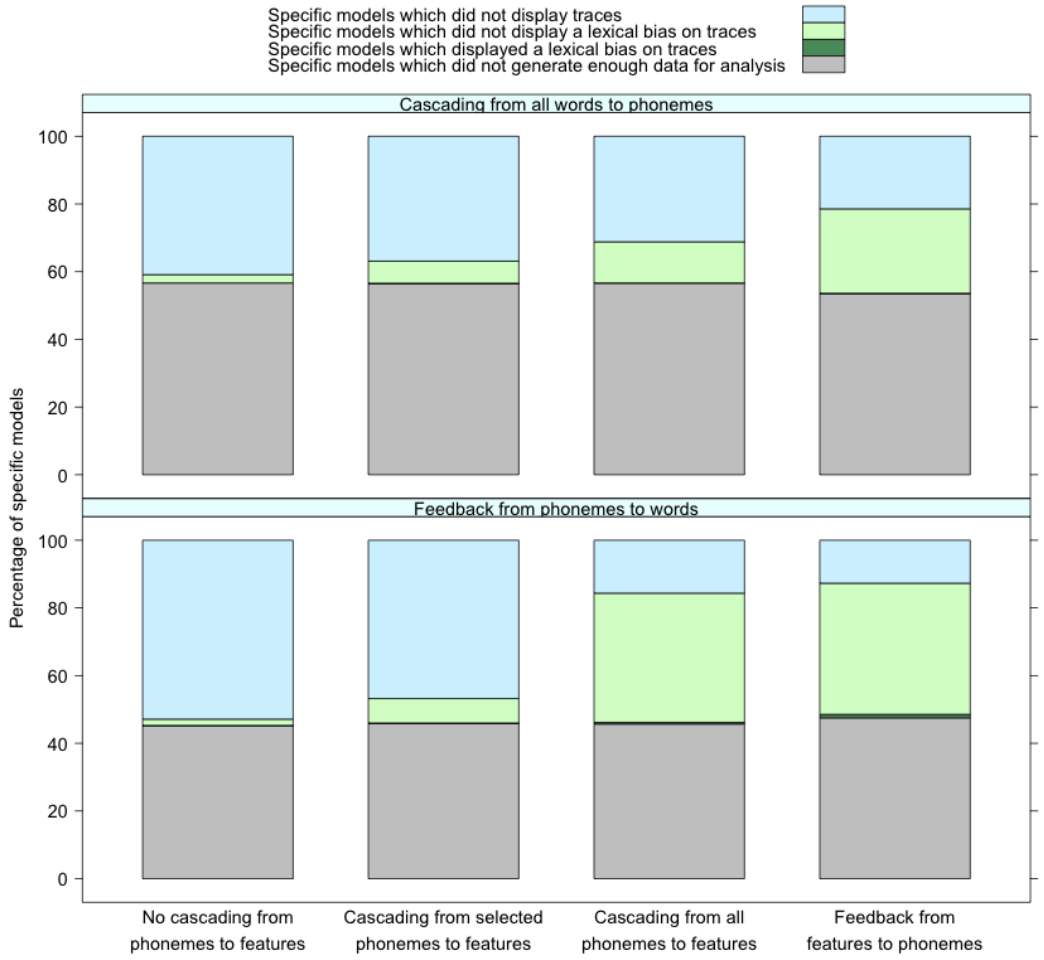


Figure 7.33: The effect of modifying activation flow on whether two-stage models generate smaller traces on lexical than on non-lexical error outcome productions of both voiceless and voiced onset consonants.

In addition, figure 7.33 and table 7.36 demonstrate that there is no evidence that any specific models with any of the architectures can simultaneously account for a lexical bias on traces on both voiced and voiceless outcomes. This may be partially due to the reduced power of the binomial analysis when analysing for the simultaneous presence of two effects, as explained in chapter 6, combined with a potentially weak underlying lexical bias on traces effect. In future work, it would be worth increasing the number of productions of each target and competitor phrase to increase power and try and address this problem.

Table 7.34: Binomial analysis to determine which two-stage architectures generate smaller traces of intended voiceless phonemes on voiced productions for lexical than for non-lexical error outcome productions, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating lexical bias effects by chance.

	Specific model counts				**Prob.**
	Total	Excluded	Sufficient data and traces	Lexical bias on traces	
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	734	38	2	0.296
Cascading from selected Ps to Fs	2916	737	109	5	0.464
Cascading from all Ps to Fs	2916	746	192	10	0.366
Feedback from Fs to Ps	5832	2299	344	19	0.276
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	2313	69	5	0.131
Cascading from selected Ps to Fs	5832	2297	234	19	0.014
Cascading from all Ps to Fs	5832	2481	521	30	0.184
Feedback from Fs to Ps	5832	2582	423	22	0.370

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

Table 7.35: Binomial analysis to determine which two-stage architectures generate smaller traces of intended voiced phonemes on voiceless productions for lexical than for non-lexical error outcome productions, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating lexical bias effects by chance.

	Specific model counts				**Prob.**	
	Total	Excluded	Sufficient data and traces	Lexical bias on traces		
Cascading from all Ws to Ps						
No cascading from Ps to Fs	2916	741	39	2	0.309	
Cascading from selected Ps to Fs	2916	741	125	6	0.435	
Cascading from all Ps to Fs	2916	742	180	9	0.413	
Feedback from Fs to Ps	5832	2097	331	15	0.591	
Feedback from Ps to Ws						
No cascading from Ps to Fs	5832	2397	72	3	0.488	
Cascading from selected Ps to Fs	5832	2389	310	30	$< .001$	*
Cascading from all Ps to Fs	5832	2600	589	57	$< .001$	*
Feedback from Fs to Ps	5832	2552	437	32	0.013	

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

Table 7.36: Binomial analysis to determine which two-stage architectures generate smaller traces of intended voiced phonemes on voiceless productions and intended voiceless phonemes on voiced productions for lexical than for non-lexical error outcome productions. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating lexical bias effects by chance.

	Specific model counts			**Prob.**
	Total	Sufficient data and traces	Lexical bias on traces	
Cascading from all Ws to Ps				
No cascading from Ps to Fs	2916	72	0	$> .9$
Cascading from selected Ps to Fs	2916	194	4	$> .9$
Cascading from all Ps to Fs	2916	356	1	$> .9$
Feedback from Fs to Ps	5832	1462	8	$> .9$
Feedback from Ps to Ws				
No cascading from Ps to Fs	5832	112	4	$> .9$
Cascading from selected Ps to Fs	5832	429	9	$> .9$
Cascading from all Ps to Fs	5832	2255	29	$> .9$
Feedback from Fs to Ps	5832	2321	61	$> .9$

Key:

Ws = words, Ps = phonemes, Fs = features

Prob. = probability

The effects of parameter manipulations on lexical bias generation on VOT traces

In the previous section, we showed that architectures with feedback from phonemes to words and cascading from selected phonemes together (therefore including the architecture with cascading from all phonemes, and the architecture with feedback from features) demonstrate a lexical bias on traces, and argued that this was due to stronger activation of unintentionally selected phonemes in the lexical error outcome condition. However, we also showed that models with feedback from phonemes to words and no cascading from phonemes can account for a lexical bias on traces, due to more trace-free phoneme errors being generated in the lexical error outcome condition.

We first examine the effect of manipulating spreading activation parameters for all architectures with feedback from phonemes to words and cascading from selected phonemes together. Results for architectures with different phoneme-to-feature activation flow options reflect the general pattern reported here. As so few specific models demonstrated a lexical bias on traces for both voiced and voiceless outcome simultaneously, we report the effects of parameter manipulations on voiced outcome errors only. Differences for voiceless errors are noted.

Again, as Goldrick and Blumstein (2006) reported lexical bias as a post-hoc test on their trace data, we only consider whether a lexical bias on traces is present on specific models which show significant trace effects. Table 7.37 and figure 7.34 show that, amongst these specific models, models with high connection strength, a high jolt to prime ratio, a low decay rate, a high number of steps before selection, and low levels of activation-based noise are most likely to display a lexical bias on traces. Results for voiceless outcome errors are similar, although the effect of decay and activation-based noise is not significant.

We showed in section 7.3 that higher numbers of steps before selection supports lexical bias generation for categorised errors, by allowing more time for activation flow. A higher connection strength increases activation flow through the feedback loops which underlie the lexical bias effect, but also increases the strength of emission of activation level signals from phoneme selection to the feature level. In a similar vein, a higher jolt to prime ratio may lead to more specific models demonstrating a lexical bias on traces by decreasing the influence of the prime on the activation level of the unintentionally selected competitor, and allowing the structure of the network and its feedback loops to have a greater effect. Finally, a low decay rate

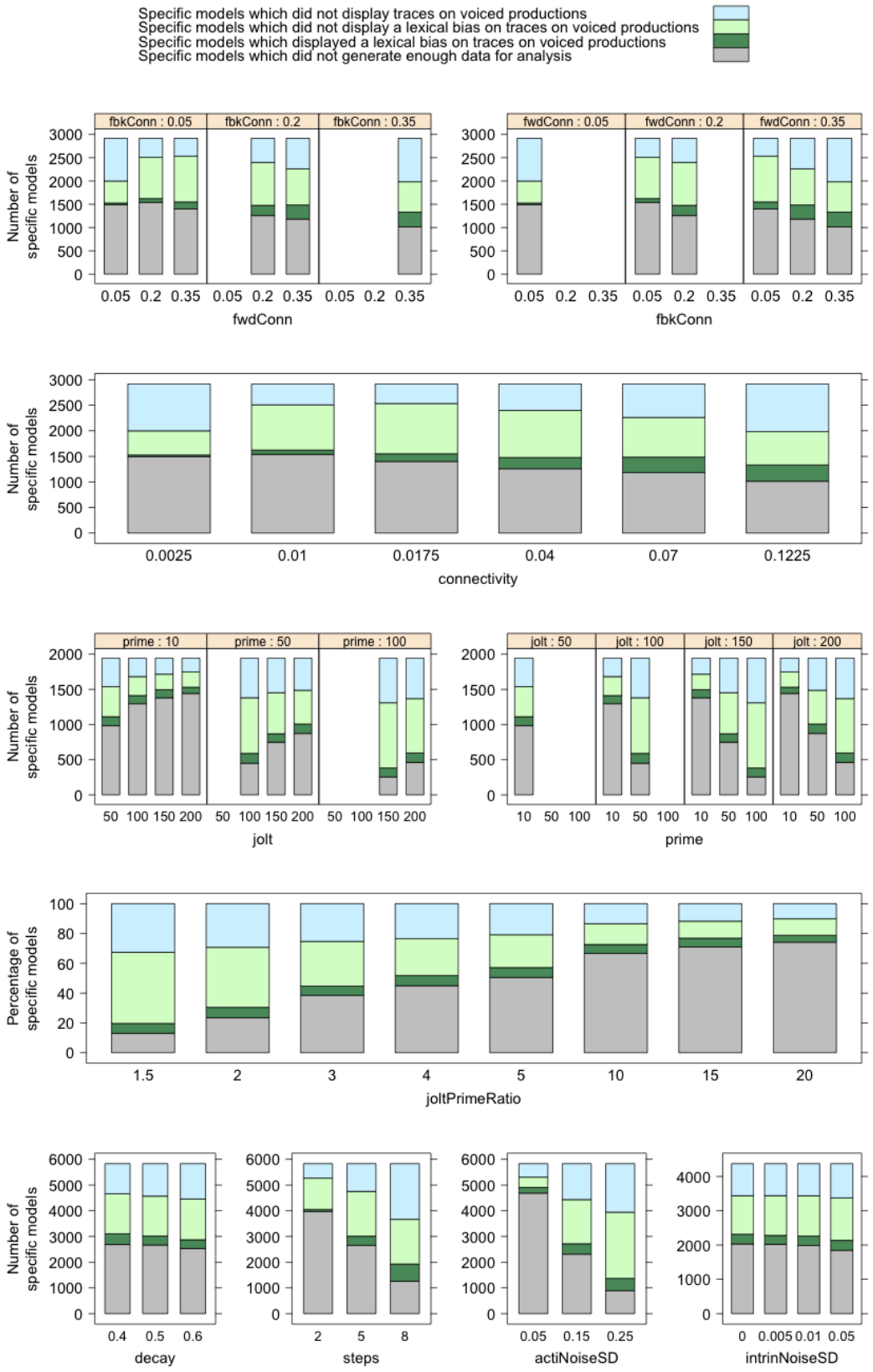


Figure 7.34: The effect of parameter manipulations on models' ability to generate traces on whether models generate smaller traces on lexical than on non-lexical error outcome productions of voiced onset consonants, in two-stage models with feedback from phonemes to words and cascading from selected phonemes.

Table 7.37: Results of logistic regression model analyses using parameter values to predict whether models generate smaller traces on lexical than on non-lexical error outcome productions of voiced onset consonants, for all two-stage models which generated traces on voiced productions with feedback from phonemes to words and cascading from selected phonemes. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	+	19.1	385	< .001	*
joltPrimeRatio	+	6.4	39	< .001	*
decay	−	6.4	41	< .001	*
steps	+	15.5	273	< .001	*
actiNoiseSD	−	3.3	11	0.001	*
intrinNoiseSD	−	0.2	0	0.859	

and a low level of activation-based noise are likely to help maintain the signal transmitted from the phoneme level, so that small differences in selection strength are still detectable at the feature level.

However, as we have previously shown, models with high connection strength and a high number of steps before selection are more likely to be ruled out by the constraints on error rate and non-contextuality of errors, and the constraint on non-contextuality of errors also causes problems for specific models with a very high jolt to prime ratio, as demonstrated in figure 7.35. As argued in the previous section, a larger number of trials per simulation would perhaps increase the power of the lexical bias on traces tests such that more models would survive this cull.

Models with feedback from phonemes to words and no cascading from phonemes to features use a different mechanism for accounting for a lexical bias on traces which relies on differences between phoneme errors and feature errors as previously explained. Figure 7.36 and table 7.38 show that in these models, the key determinant of whether a lexical bias is detected is the amount of intrinsic noise in the network. This is likely to reflect that in models where more feature errors are generated upon which traces are evident, the difference between the lexical and non-lexical conditions will be clearer. However, extremely noisy models are ruled out by the constraints on error rate and non-contextuality of errors, as shown in figure 7.37.

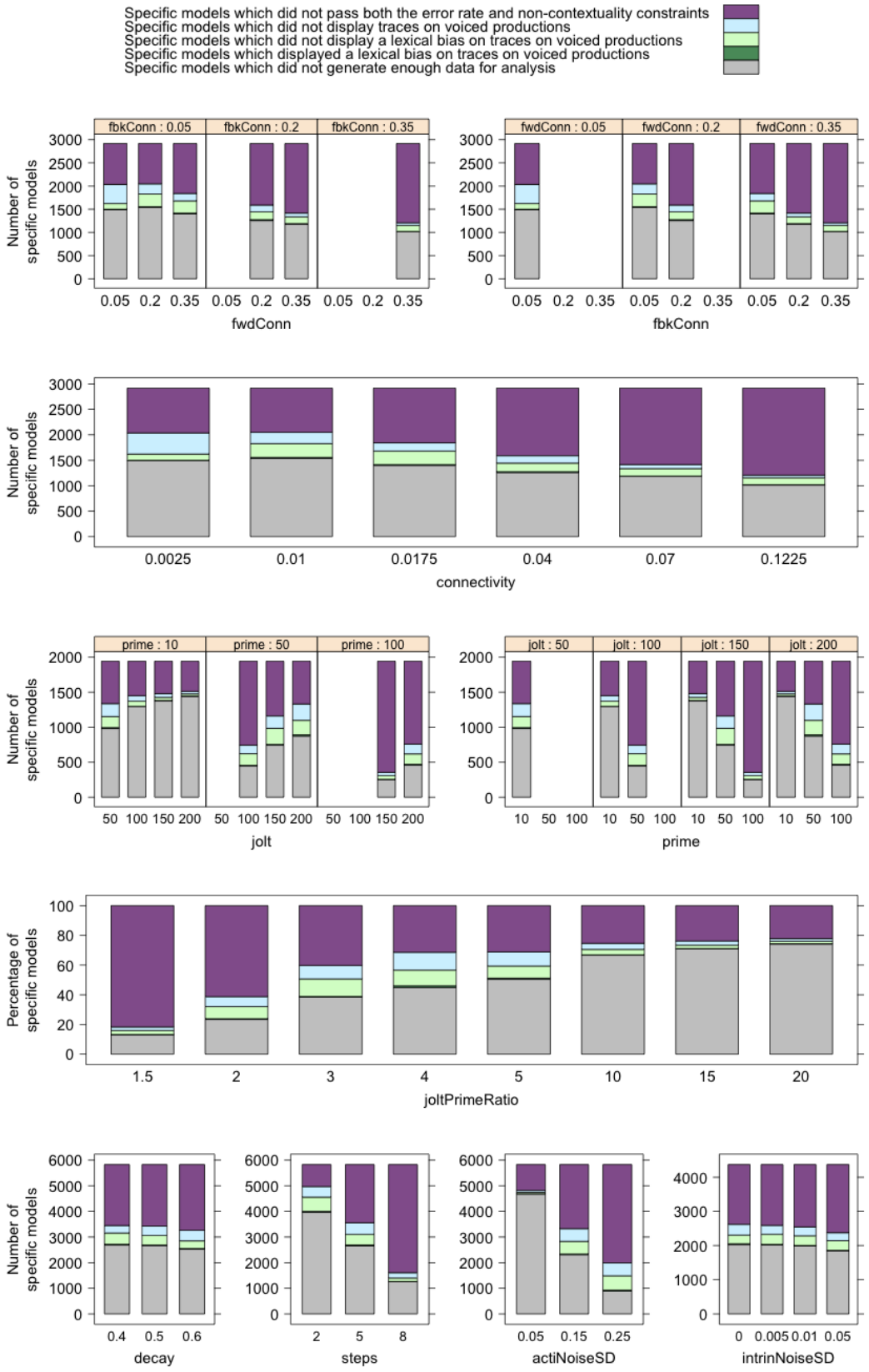


Figure 7.35: The effect of parameter manipulations on models' ability to generate traces on whether models generate smaller traces on lexical than on non-lexical error outcome productions of voiced onset consonants, in two-stage models with feedback from phonemes to words and cascading from selected phonemes, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

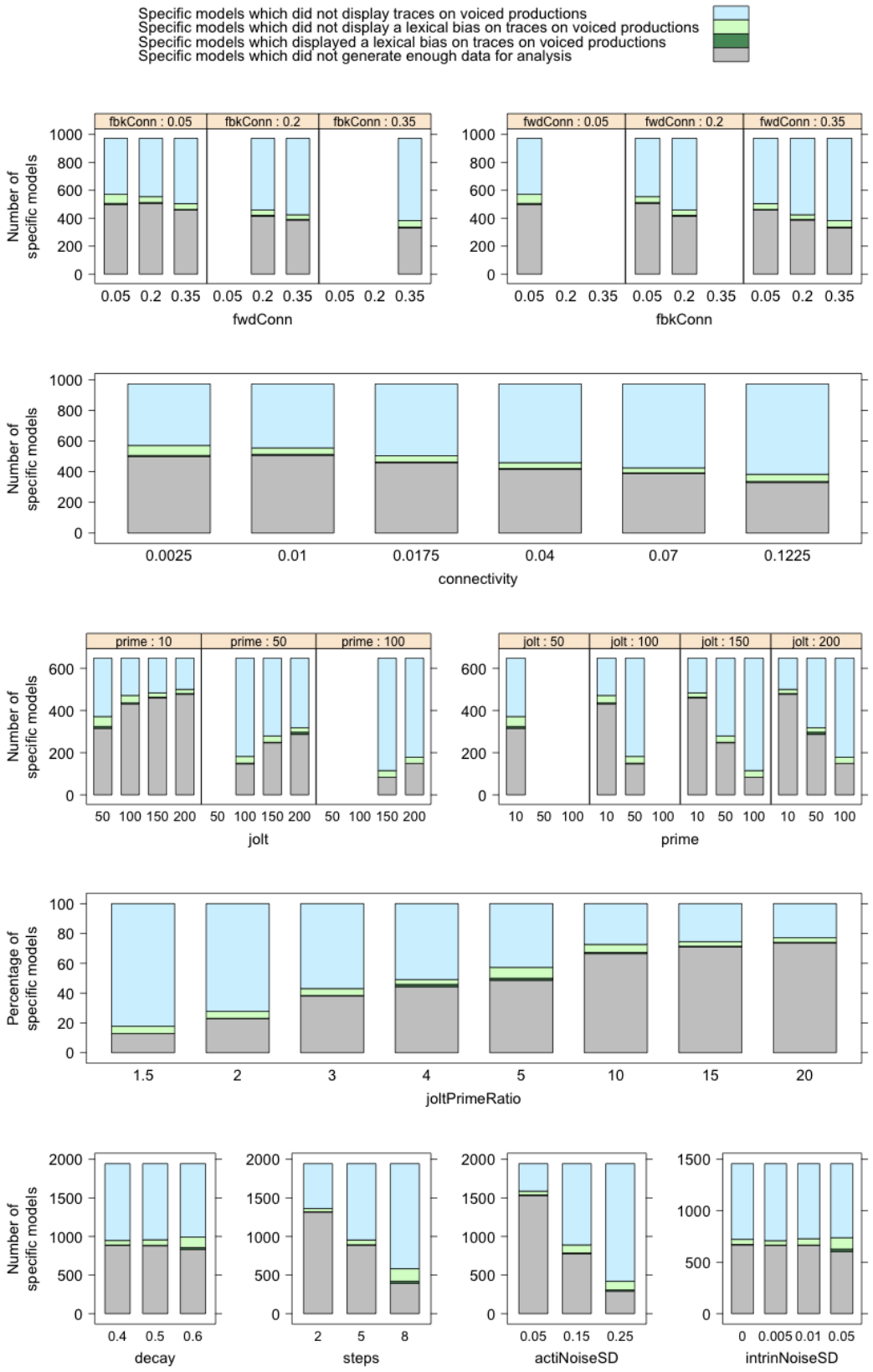


Figure 7.36: The effect of parameter manipulations on models' ability to generate traces on whether models generate smaller traces on lexical error outcome productions than non-lexical error outcome productions of voiced onset consonants, in two-stage models with feedback from phonemes to words and no cascading from phonemes.

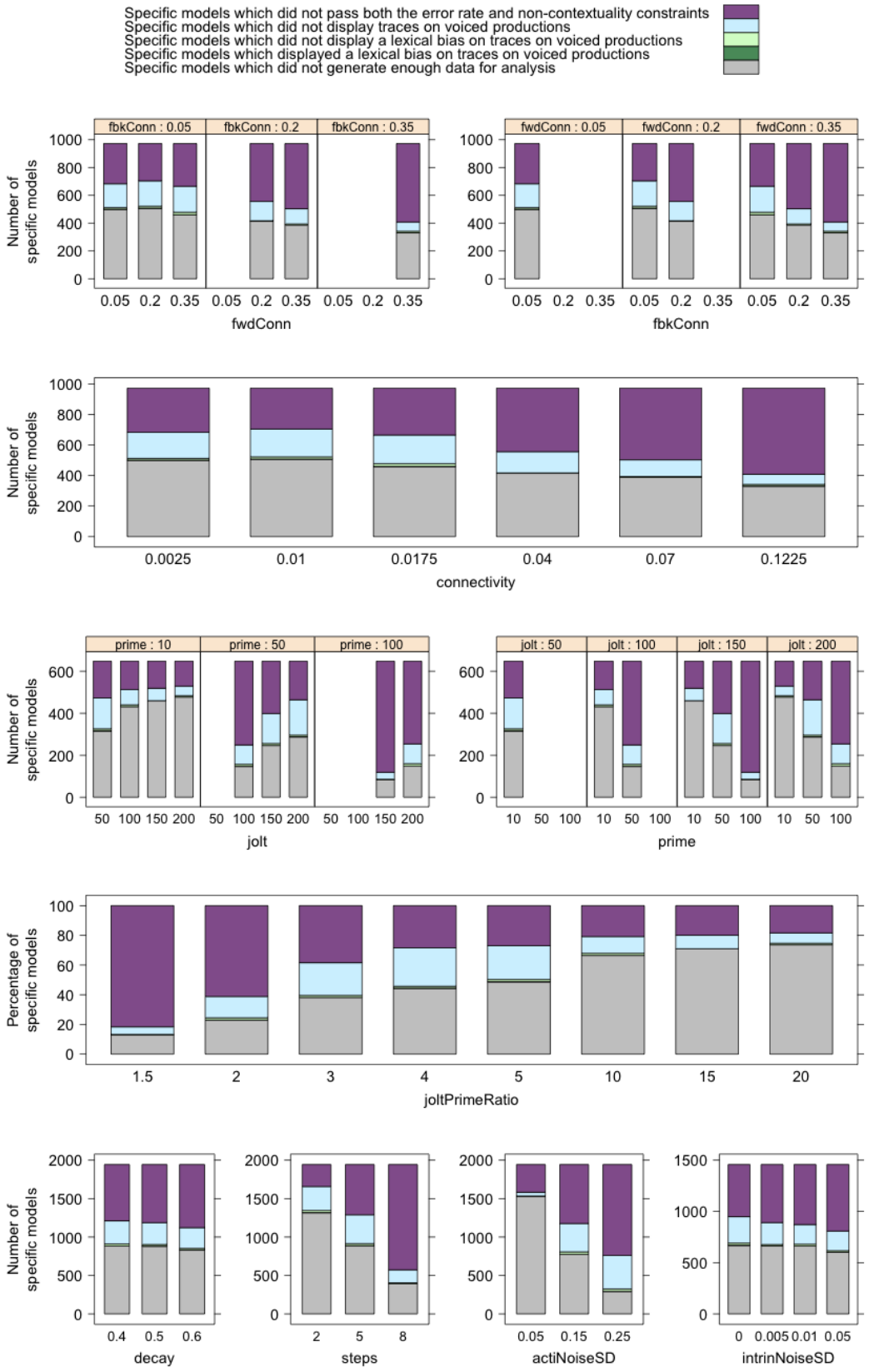


Figure 7.37: The effect of parameter manipulations on models' ability to generate traces on whether models generate smaller traces on lexical error outcome productions than non-lexical error outcome productions of voiced onset consonants, in two-stage models with feedback from phonemes to words and no cascading from phonemes, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

Table 7.38: Results of logistic regression model analyses using parameter values to predict whether models generate smaller traces on lexical error outcome productions than non-lexical error outcome productions of voiced onset consonants, for all two-stage models which generated traces on voiced productions with feedback from phonemes to words and no cascading from phonemes. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	+	1.0	1	0.310	
joltPrimeRatio	+	0.1	0	0.893	
decay	+	0.5	0	0.639	
steps	+	1.3	2	0.175	
actiNoiseSD	+	0.7	0	0.490	
intrinNoiseSD	+	2.3	6	0.017	*

7.5.3 Conclusions

In this section, we replicated our findings from section 7.4 that cascading from phonemes to features is not required for traces of intended phonemes to be found on erroneous productions, using a bigger lexicon and different materials. Against both Goldrick and Blumstein's (2006) and our own predictions, we also found that cascading from phonemes to features is not required for specific models to demonstrate smaller traces in the lexical error outcome condition in comparison to the non-lexical error outcome condition. We argue that the architecture with no cascading from phonemes exhibits this effect because more phoneme level errors occur in the lexical error outcome condition than the non-lexical error outcome condition, whilst feature level errors are not affected by this variable. As traces only occur on errors at the feature level, and there is a higher proportion of errors at the phoneme level in the lexical error outcome condition in comparison to the non-lexical error outcome condition, traces are on average smaller in the lexical error outcome condition.

Over all architectures however, few models show this effect, probably because it is quite a weak effect. Models which generate enough errors for the effect to be detected get ruled out when constraints on error rate and non-contextuality of errors are applied. The effect is also not strong enough to be detected by the lower power binomial analysis examining which architectures demonstrate lexical bias on traces on both voiced and voiceless outcome errors simultaneously. Future investigations could increase the power of tests on specific models by increasing the number of

trials per simulation. Alternatively, it would be interesting to investigate whether the size of the effect generated by the model can be compared to the size of the effect in humans. Evidence that the effect generated in the current implementation is too weak to represent a good model of the effect in humans would constitute motivation to consider modifications to the architecture to strengthen the effect.

An analysis of the effects of parameter manipulations on whether models demonstrate a lexical bias on traces showed that in architectures with cascading from selected phonemes, models with a higher number of steps before selection are more likely to display this effect because such parameter settings allow activation to flow through feedback loops for longer. Models with higher connection strength also support the effect by boosting feedback loop activation, but also by strengthening the signal transmitted from phoneme selection to the featural level. Models with higher jolt to prime ratios are more likely to show this effect because higher jolt to prime ratios lead to a weaker influence of the prime activation on the activation level of a misselected phoneme, so that this level reflects the structure of the network and its feedback loops more strongly. Finally, low decay rates and low levels of activation-based noise aid accurate transmission of activation output from the phoneme level to the feature level. In architectures with no cascading from phonemes, models with high levels of intrinsic noise are more likely to exhibit this effect as trace effects are stronger in these models due to more featural errors occurring. In these models it is therefore easier to detect an effect of lexicality on VOT traces.

7.6 Conclusions

In this section, we introduced a two-stage phonological encoding and subphonemic processing model with output at the featural level. We showed that a model with no cascading from phonemes to features can account for both classic results from transcribed records of speech errors, and new evidence where acoustic properties of correct and erroneous productions are compared. The studies reported here represent the first time VOT evidence has been modelled in simulations based on Dell's (1986) model.

Specifically, we found that not only can a model with no cascading from phonemes to features account for the classic lexical bias effect, but it can also explain the classic phonological similarity effect. No feedback is required to account for the phonological similarity effect in this model, as misselections of one feature are more likely than misselections of two. Furthermore, our simulations of Goldrick and

Blumstein's (2006) experiments showed that errors at the featural level result in traces of intended phonemes on errors when their VOTs are compared to those from correct productions, such that cascading from phonemes is not required to account for this effect either, contrary to Goldrick and Blumstein's (2006) claims. Lastly, we found that contrary to our own predictions as well as Goldrick and Blumstein's (2006), a model with no cascading from phonemes to features can also explain smaller VOT traces in the lexical error outcome condition. In this architecture, when feedback from phonemes to words is present, more errors are generated at the phoneme level but not the feature level in the lexical error outcome condition. However, traces are present on feature errors only, such that a greater number of phoneme errors results in smaller average trace size.

We noted that in the current implementation where priming is applied at the word level, prime activation can only cause contextual error generation at the featural level when activation cascades from all phonemes and feedback loops are present to reinforce the prime activation so that it does not decay during phonological encoding. In other architectures, high proportions of non-contextual errors are therefore generated at the featural level. Where an architecture's account of an effect relies on featural error generation but there is no priming support for these errors, specific models tend to either display too weak an effect for the statistical test to detect it, due to paucity of data; or sufficient data and a significant effect, but correspondingly a very high overall error rate. These specific models are therefore often excluded when the constraints on error rate and non-contextuality of errors are applied. Similarly, as so few models generate sufficient errors, problems arise when utilising binomial analyses to confirm whether an architecture can account for multiple effects simultaneously, as our binomial analyses for multiple effects have lower power as explained in chapter 6.

When simulating the phonological similarity effect, this issue causes problems for all architectures without feedback from features to phonemes apart from the architecture with feedback from phonemes to words and cascading from all phonemes; and when simulating traces of intended phonemes on unintended phonemes, all architectures with no cascading from phonemes experience these difficulties. Future work will seek to verify that applying priming at the featural level removes this problem. Pending such work however, we are still able to demonstrate that models with no feedback from features to phonemes which exhibit appropriate error rates and proportions of non-contextual errors can account for the phonological similarity effect, as contextual errors at the featural error are primed in the architecture with

feedback from phonemes to words and cascading from all phonemes. Furthermore, models with cascading from selected phonemes only which exhibit appropriate error rates and proportions of non-contextual errors can account for VOT traces of intended phonemes on errors, as the voicing characteristics of intended productions are more emphasised due to stronger activation of intentionally selected phonemes compared to unintentionally selected phonemes.

We also found that the lexical bias effect on traces exhibited in all of our architectures with feedback from phonemes to words is weak. As a result, the effect is only detected in specific models with high error rates, and these specific models are ruled out when the constraints on error rate and non-contextuality of errors are applied. Similarly, problems are experienced when a binomial analysis is used to investigate whether lexical bias effects are generated on traces for both voiced outcome productions and voiceless outcome productions simultaneously due to the lower power of the multiple effects binomial analysis. Future work could address this by increasing the number of trials per simulation. Alternatively, evidence that the effect generated in the current implementation is too weak to represent a good model of the effect in humans would constitute motivation to consider modifications to the architecture to strengthen the effect.

Throughout this chapter, we clarified which parameter settings are required for architectures to be able to account for certain effects. We found that for architectures which relied on feedback loops to generate the lexical bias and phonological similarity effects, specific models with high connection strengths, a high number of steps before selection and high levels of activation-based noise were most successful, as these increased the influence of the feedback loops. Models which successfully generated phonological similarity effects using errors at the featural level tended to have parameter settings previously shown to cause high error rates: i.e., low forward connection strength, high levels of decay, high numbers of steps before selection and high levels of intrinsic noise.

These same error inducing parameter settings were shown to enable models to exhibit traces of intended phonemes on VOT measurements using errors at the featural level. It was also shown that a high jolt to prime ratio was useful as this reduces the number of phoneme level errors generated. Traces originating in misselection at the phoneme level however were best supported by high forward connection strengths, low decay rates and low numbers of steps before selection, as these parameter settings allowed the activation patterns created at phonological encoding to be faithfully transmitted to the feature level without becoming distorted. Models in

which phoneme level contextual error rates were boosted by low jolt to prime ratios and high levels of activation-based noises were also more likely to show significant traces.

There are clear differences in the parameter settings required for trace effects to be displayed in different architectures. For example, a high level of decay and high numbers of steps before selection support trace generation in architectures with no cascading from phonemes, whereas a low level of decay and a low number of steps support trace generation in other architectures. This result demonstrates that our concern that testing different architectures at one arbitrary set of parameter settings may lead to misleading results was not unfounded.

We found that the parameter settings required for lexical bias and phonological similarity effects were not incompatible with the parameter settings required for trace generation, such that a binomial analysis showed that architectures with feedback from phonemes to words and either cascading from all phonemes or feedback from features to phonemes could account for all effects simultaneously. It is difficult to rule out the possibility that other architectures (e.g., the architectures with feedback from phonemes to words but either no cascading from phonemes or cascading from selected phonemes only) can account for these effects, as the phonological similarity effect in these architectures is weak due to our word level priming implementation decision, and the power of the binomial analysis when considering multiple simultaneous effects is reduced. Similar concerns about power apply to our finding that no architecture can account for all effects when the constraints on error rate and non-contextuality of errors are applied. This highlights a need to improve the binomial analysis of multiple simultaneous effects in future work.

Finally, we found that in architectures with cascading from selected phonemes, a lexical bias on VOT traces of intended phonemes was best supported by parameters which boosted the lexical bias effect (high numbers of steps before selection, high connection strengths, high jolt to prime ratios) and supported transmission of lexical bias activation patterns to the feature level (high connection strengths, high jolt to prime ratios, low decay rates and low levels of activation-based noise). Architectures with no cascading from phonemes displayed were more likely to display a lexical bias on traces when levels of intrinsic noise were high. In these models, more errors occur at the featural level, which supports trace generation such that an effect of lexicality on traces would be easier to detect.

7.7 Chapter summary

In this chapter, we introduced a two-stage model of phonological encoding and subphonemic processing, with output at the featural level. In line with our predictions, we demonstrated that in a two-stage model, cascading from phonemes is not required to account for the classic lexical bias effect, the classic phonological similarity effect (unlike in a one-stage model), or VOT traces of intended phonemes on unintended productions, in contrast to claims made by Goldrick and Blumstein (2006). We found that different parameter settings were required for different accounts of trace generation, demonstrating that a test of the ability of different architectures to account for this evidence at one set of arbitrary parameter settings would not have been a fair test. Contrary to our own predictions as well as Goldrick and Blumstein's (2006), we further showed that cascading from phonemes is not required to explain lexical bias effects on VOT traces (Goldrick & Blumstein, 2006).

As well as showing that these empirical results do not constrain models of activation flow between phonemes and features, these simulations demonstrate for the first time that instrumental evidence of word production can be modelled within the framework of Dell's (1986) architecture.

Chapter 8

Activation flow between phonemes and features: instrumental evidence abandoning categorisation

8.1 Introduction

In the previous chapter, we began to consider a two-stage phonological encoding and subphonemic processing model, and used simulations to examine the constraints placed on models of activation flow between phonemes and features by evidence which relies on the categorisation of productions as erroneous or correct. We found that no cascading from phonemes or feedback from features was required to explain any of the effects we modelled, including the lexical bias effect as reported in transcribed records of speech errors (e.g., Dell & Reich, 1981; Hartsuiker et al., 2005), the phonological similarity effect as reported in transcribed records of speech errors (e.g., Levitt & Healy, 1985; Nooteboom, 1969), VOT traces of intended phonemes on erroneous productions as reported by Goldrick and Blumstein (2006), and lexical bias effects on VOT traces of intended phonemes as reported by Goldrick and Blumstein (2006). These investigations also constituted the first simulations of acoustic VOT evidence in the framework of Dell's (1986) model.

In this chapter, we aim to extend these studies in two ways. Firstly, we build on the successful simulations of VOT evidence in the previous chapter with simulations of electropalatography (EPG) and ultrasound evidence. Secondly, we begin to consider findings which do not rely on categorisation of productions as erroneous or correct. We investigate whether these results constrain models of activation flow between phonemes and features any further.

In order to make it possible to draw conclusions from instrumental data without categorising productions as erroneous or correct, McMillan et al. (2009) present the *delta method*. The delta method permits a similarity value to be calculated for

two measurements of articulation. As explained in section 2.3.2, in an experiment where materials were manipulated so that for half the materials, onset errors would result in words, and for the other half, onset errors would result in non-words, McMillan et al. (2009) showed that in the lexical condition articulations of onset phonemes are significantly more like reference measurements for the competing place of articulation than they are in the non-lexical error outcome condition. However, no significant difference between the two conditions was found for similarity of articulations to the reference measurement for the target place of articulation.

McMillan et al. (2009) presented this finding as evidence for feedback from phonemes to words. They further noted that a model in which activation cascades from unselected phonemes would predict this result, as extra activation conveyed to the competing phoneme in the lexical outcome condition would cascade to the feature layer. However, we observed that these results can be explained in any model of information flow from phonological encoding to subphonemic processes, as long as feedback from phonemes to words is present. Even in a model with no cascading from phonological encoding, more frequent production of the competing onset in the lexical condition would lead to an average articulation closer to the reference measurement for the competing place of articulation than the average articulation in the non-lexical condition.

We first simulate McMillan et al.'s (2009) results, to demonstrate that the delta method can be used to evaluate simulation output, to show that EPG evidence can be simulated within Dell's (1986) model, and to verify that all architectures with feedback from phonemes to words exhibit this effect. We focus on determining which architectures can account for the significant difference found for comparisons to the reference competitor place of articulation, as without knowledge of the power of McMillan et al.'s (2009) experiment, we do not know how reliable the finding of no effect of lexicality on the similarity of articulations to a target competitor was.

We then consider McMillan's (2008) finding of phonological similarity effects on articulations. McMillan (2008) applied the delta method introduced by McMillan et al.'s (2009) to analyse EPG, ultrasound and VOT measurements. As described in section 2.3.2, the articulatory results showed that articulatory measurements were further from a reference measurement for the target onset when target and competing onsets differed in place than when the onsets differed in both place and voicing. Similarly, the acoustic results showed that acoustic measurements were further from the reference when the onsets differed in voicing than when the onsets differed in both place and voicing (although this result failed to reach significance in

the ultrasound study). McMillan (2008) argued that this was evidence for feedback from features to phonemes. Competing phonemic representations which share more features (and therefore differ by fewer features) will receive more activation via feedback from subphonemic representations. Activated phonemic representations will then pass activation to their component subphonemic representations, including the competing voicing or place representation. Because similar phonemes receive more activation from the target phoneme, the competing voicing or place subphonemic representation will receive more activation when a more similar phoneme is competing.

The results simulated so far in this thesis, including the transcribed phonological similarity effect, have presented no constraints on models of activation flow between phonemes and features. We further predict that McMillan et al.'s (2009) results will not constrain these models either. Simulation results demonstrating that an account of McMillan's (2008) results does require feedback from features to phonemes would therefore place particular importance on McMillan's (2008) findings.

8.2 McMillan et al.'s (2009) evidence of a lexical bias on articulatory measurements

In this section, we investigate whether we can simulate McMillan et al.'s (2009) findings that articulations of onset phonemes in a lexical error outcome condition are significantly more like reference articulation measurements for the competing place of articulation than they are in a non-lexical error outcome condition, and clarify the constraints this result places on model architecture and parameter settings. This constitutes the first attempt to model EPG results within Dell's (1986) architecture. We predict that feedback from phonemes to words will be required to account for this result, but no cascading from phonemes to features will be necessary.

8.2.1 Simulation methodology

To investigate which architectures could account for McMillan et al.'s (2009) findings, we carried out a delta analysis of the output of the lexical bias simulations described in section 7.3.

Model configuration

All 37,908 two-stage models were tested.

Model task and lexicon

As described in section 7.3, the model produced single words while competitor words were primed. The target materials and 100 word lexicon are described in full in section 6.2.2. Each material set contains 16 target and competitor combinations, and each of these target words (along with the corresponding competitor) was produced 500 times by each specific model, resulting in a total of 8000 word productions. There were in fact a number of extra productions at the beginning of each simulation, for the purpose of calculating reference utterances. On these productions, the competitor for the target word was itself, such that the target word received both jolt and prime activation, to simulate a situation in which no onset error was induced. Each of the eight target words in the material set in use was produced in this manner 500 times. In this study, we consider behaviour from the simulation using the material set in which place of articulation of the target and competitor onset always differs, to allow us to simulate McMillan et al.'s (2009) EPG experiment.

Model output interpretation

When using the delta method, it is not necessary to categorise productions as correct or erroneous. Instead, to compare the behaviour of the model to the human behaviour reported by McMillan et al. (2009), we recorded the activation of the alveolar and velar features, which we take to abstractly represent the extent to which the resulting articulation involves tongue raising at the front and the back of the mouth respectively, as explained in chapter 3.

Output from the productions in which a target word received both jolt and prime activation was used to calculate an alveolar onset reference utterance vector and a velar onset reference utterance vector. In each case this involved determining the average alveolar feature activation and velar feature activation of intended alveolar or velar productions.

In the main part of the simulation, for each production of the target words with primed competitors, the Euclidean distance between the alveolar and velar feature activation vector recorded from that production and the reference utterance vector for the competing place of articulation was calculated, to represent delta. A t-test comparison of delta measurements recorded in the lexical outcome condition and delta measurements recorded in the non-lexical outcome condition was carried out. This allowed us to determine which specific models demonstrated a smaller delta

Table 8.1: Binomial analysis to determine which two-stage architectures display a smaller delta measured from the competitor reference for stimuli with a lexical error outcome, for vectors of alveolar and velar feature activation values used to simulate tongue-to-palate contact. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating lexical bias effects by chance.

	Specific model counts			**Prob.**
	Total	Sufficient data	Lexical bias on delta	
Cascading from all Ws to Ps				
No cascading from Ps to Fs	2916	2916	119	$> .9$
Cascading from selected Ps to Fs	2916	2916	109	$> .9$
Cascading from all Ps to Fs	2916	2916	115	$> .9$
Feedback from Fs to Ps	5832	5832	87	$> .9$
Feedback from Ps to Ws				
No cascading from Ps to Fs	5832	5832	217	$> .9$
Cascading from selected Ps to Fs	5832	5832	83	$> .9$
Cascading from all Ps to Fs	5832	5832	93	$> .9$
Feedback from Fs to Ps	5832	5832	69	$> .9$

Key:
Ws = words, Ps = phonemes, Fs = features
Prob. = probability

from the competitor in the lexical outcome condition, as reported by McMillan et al. (2009). A similar t-test comparison was carried out for delta values calculated from the reference utterance vector for the target place of articulation, for information only.

8.2.2 Simulation results

We begin our analysis by investigating which architectures can account for McMillan et al.'s (2009) results.

Architecture analysis of lexical bias effects

Table 8.1 demonstrates that, somewhat surprisingly, there is no evidence that any of the architectures can account for this empirical finding. This result is not affected by excluding specific models which generate no errors according to categorisation, or by excluding specific models which fail either the constraint on error rate or the constraint on non-contextuality of errors. We note for information that there is no evidence for any lexicality effect on calculations of distance from the reference articulation for the target place of articulation either.

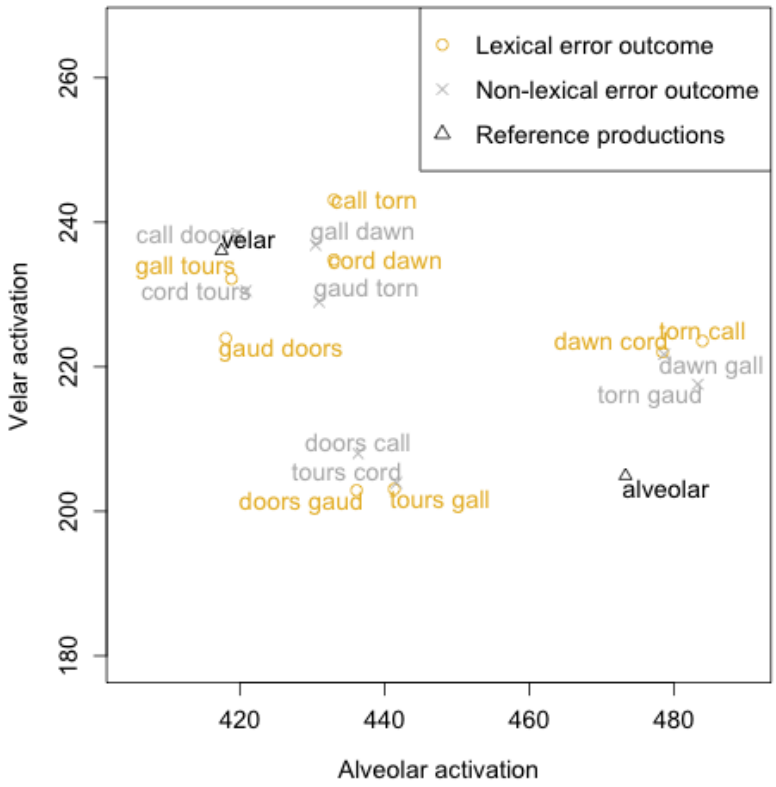


Figure 8.1: Mean alveolar and velar activation values for productions of single words when competitor words are primed, in two-stage models with feedback from phonemes to words. The same distance is used to represent one unit of activation on the x-axis as on the y-axis.

Table 8.2: Frequency of occurrence of stimulus onsets and codas in the model's lexicon.

Onsets		**Codas**	
/k/	6	/n/	13
/g/	5	/l/	9
/t/	5	/d/	8
/d/	4	/z/	7

Alveolar and velar feature activation

To try and understand why none of the architectures could capture this effect, we looked at the average alveolar and velar feature activation values being generated for each target word and competitor pair. We considered output from models with feedback from phonemes to words only, as we expected that this feedback was required for the effect to be found.

Figure 8.1 shows that no clear effect of lexicality can be seen on the distribution of these average activation values. Instead, comparison of the diagram with the frequency statistics given in table 8.2 shows that the frequency with which onsets and codas occurs in words in the lexicon is a much stronger predictor of the activation of the alveolar and velar features in the lexicon, where more frequent onsets and codas tend to lead to more alveolar and velar activation. For example, for intended velar productions, target words with the onset /k/, which occurs as an onset 6 times in the lexicon, generally lead to more velar and also alveolar activation than target words with the onset /g/, which occurs as an onset 5 times in the lexicon. This effect can also be see when comparing intended alveolar productions with competitor words with the onset /k/ and with the onset /g/. Codas also have an effect, as is particularly noticeable when comparing the alveolar and also velar activation for alveolar target words with the coda /n/, which occurs as a coda 13 times in the lexicon, with alveolar target words with the coda /z/, which occurs only 7 times in the lexicon. The strong activation increase caused by use of the coda /n/ can also be seen where /n/ and /z/ are codas on competitor words for velar productions.

Figure 8.1 also shows that both intended alveolar and intended velar productions have much more alveolar activation than velar activation. This is because there are seven onset phonemes with alveolar place of articulation, in comparison to only two with a velar place of articulation. Where parameter settings allow feedback loops to have a large effect on the output of the model, the alveolar feature will receive activation from more phonemes than the velar feature will in architectures with cascading from all phonemes to features, and in architectures with feedback from features to phonemes, the feedback loops between the alveolar feature and the seven alveolar onset features will result in an even stronger increase of the alveolar activation level. In these models where the alveolar activation becomes higher than the velar activation due to frequency, productions will frequently be categorised as alveolar rather than velar, regardless of whether a velar or alveolar production was intended, resulting in a high error rate.

If we exclude specific models which generate too many errors or too many non-contextual errors for the constraints we established in chapter 4, productions become much more clearly classified as velar or alveolar, as can be see in figure 8.2. However, a closer look at intended alveolar productions, as in figure 8.3, shows that effects of onset and coda are still present and stronger than any potential effects of lexicality. As a side note, it is clear that in these models, the effect of the frequency of the competitor onset and coda is greatly diminished, probably because specific models

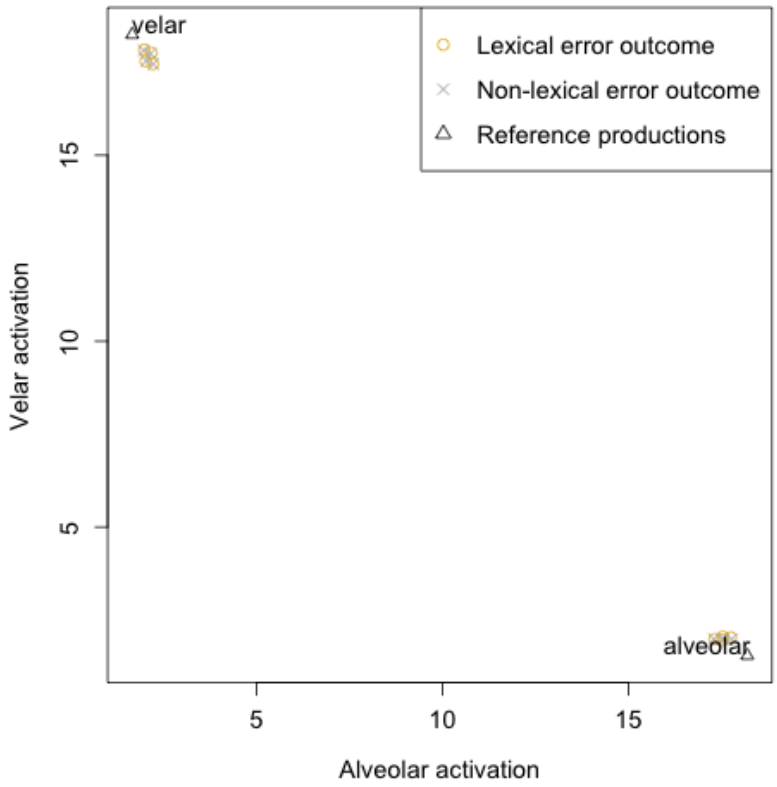


Figure 8.2: Mean alveolar and velar activation values for productions of single words when competitor words are primed, in two-stage models with feedback from phonemes to words, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. The same distance is used to represent one unit of activation on the x-axis as on the y-axis.

with relatively high primes (or in other words very low jolt to prime ratios) are ruled out for generating too many errors, as we showed in chapter 4.

We note however that table 8.1 shows that, of the architectures with feedback from phonemes to words, there are many more models with no cascading from phonemes to features displaying significant effects of lexical bias on delta. It seems reasonable to hypothesise that these models will be substantially less affected by frequency effects originating in feedback loops between phonemes and words, as activation does not cascade from phoneme selection. Figure 8.4 displays average alveolar and velar activation values for intended alveolar productions in the architecture with no cascading from phonemes, and shows that, in this architecture, there is indeed a clear effect of lexicality, such that productions where an error outcome would be lexical are further away from the alveolar reference, and presumably closer to

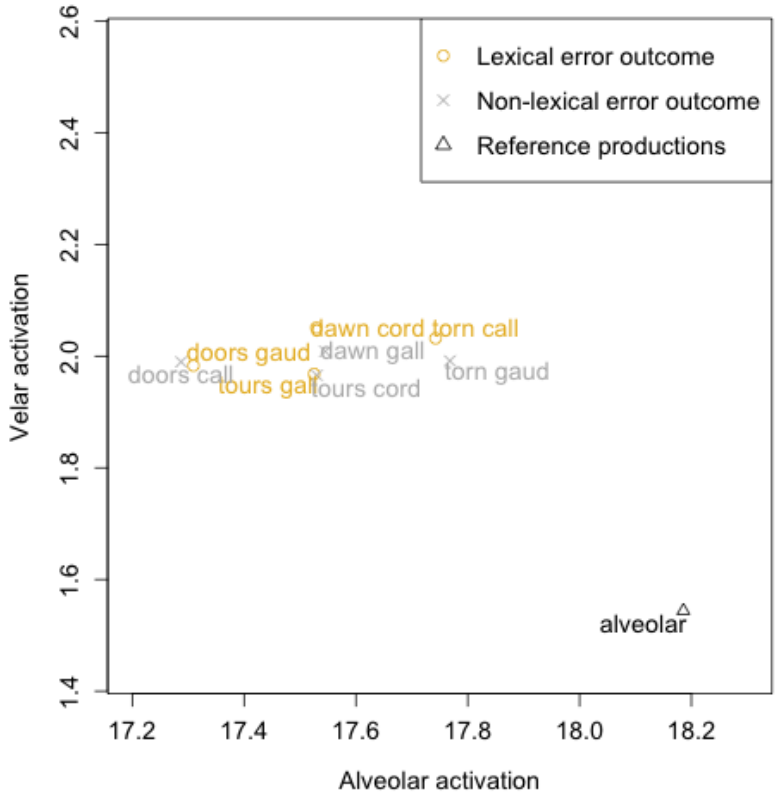


Figure 8.3: Mean alveolar and velar activation values for productions of single words with alveolar onsets when competitor words are primed, in two-stage models with feedback from phonemes to words, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. The same distance is used to represent one unit of activation on the x-axis as on the y-axis.

the velar reference (as shown by McMillan et al., 2009). However, there is still an effect of onset frequency, simply because phonemes which are more frequent will be selected more often at the phoneme level, and this effect is stronger than the lexicality effect. For example, all the target productions with a /t/ onset are more alveolar and less velar than the productions with a /d/ onset.

We conclude therefore that onset and coda frequency effects are clearly causing very large within condition variance. This makes it difficult to detect any effect of lexicality using the between-items t-test on deltas from the competitor that we carried out per specific model at simulation run time. This frequency driven within condition variance is likely to also explain why table 8.1 shows that there are actually numerically fewer models displaying a statistically significant result once feedback is added to the model. However, at least in the case of the architecture with no

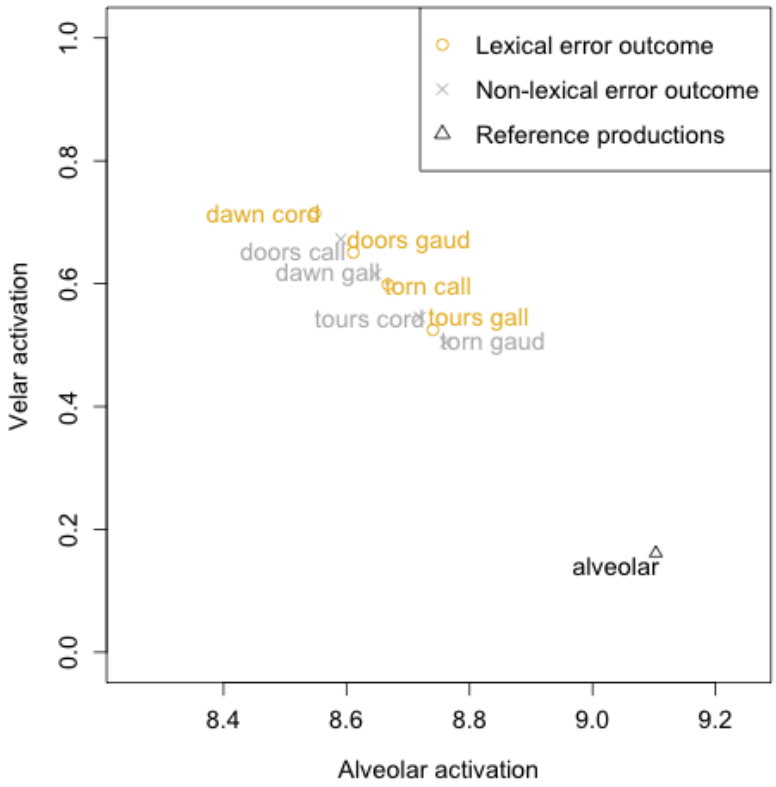


Figure 8.4: Mean alveolar and velar activation values for productions of single words with alveolar onsets when competitor words are primed, in two-stage models with feedback from phonemes to words with no cascading from phonemes to features. The same distance is used to represent one unit of activation on the x-axis as on the y-axis.

cascading from phoneme selection, we have seen evidence that there may still be an underlying trend in the right direction.

With the knowledge that frequency causes such huge variance, we could in future consider using a more complicated statistical test than the t-test to account for the influence of frequency directly. However, in this thesis, we take a first step by making a simple extension of the binomial analysis to investigate whether specific models of certain architectures really are displaying a trend in the right direction. Instead of determining the probability of finding the reported number of statistically significant results given the probability of a Type I error as we have done previously, we consider the probability of finding the reported number of specific models displaying numerical results which are in the predicted direction, given a 0.5 probability that we would find a result in the right direction by chance.

Architecture analysis of lexical bias trends

When running binomial analyses of the number of models generating statistically significant effects, we have only been interested in the probability of witnessing the same number or more specific models generating significant effects by chance. These binomial analyses are therefore one way tests, as the probability of witnessing less specific models generating significant effects by chance is a meaningless statistic. This is not the case for trends however, as the absence of a numerical difference in the right direction generally indicates the presence of a numerical difference in the wrong direction (bar the presumably few cases where results for the lexical and non-lexical conditions are numerically identical). In these analyses we therefore use two way tests, and investigate for which architectures there would be less than 0.025 probability (Bonferroni corrected to 0.003125) of witnessing the same number or more specific models generating lexical bias trends by chance.

Table 8.3 shows that, as we predicted, all architectures with feedback from phonemes to words have a higher than chance level of specific models displaying numerically smaller deltas from the reference measurement for the competing place of articulation in the lexical outcome condition in comparison to the non-lexical outcome condition. Unexpectedly, there appears to be evidence that the architecture with no feedback from phonemes to words and cascading from all phonemes to features can account for the effect. There is no clear reason why this result should be significant. Replication of this study would help establish whether this is a Type I error.

Figure 8.5 shows the number of models which show a numerical effect in the right direction. For information, we also evaluate whether architectures with feedback from phonemes to words have more specific models than would be predicted by chance displaying numerically larger deltas from the reference measurement for the target place of articulation in the lexical outcome condition in comparison to the non-lexical outcome condition. Table 8.4 shows that they do.

Our findings demonstrate that there is an effect of lexicality on delta measured from the competitor place in all architectures with feedback from phonemes to words. We leave consideration of the effect of excluding specific models which fail the constraints on error rate and non-contextuality of errors to a later date when we have demonstrated that individual specific models can demonstrate a statistically significant effect of lexicality on delta. Similarly, an analysis of which parameter settings allow the implementation to capture both this effect and others which we have modelled would be more usefully carried out when we are more sure of which

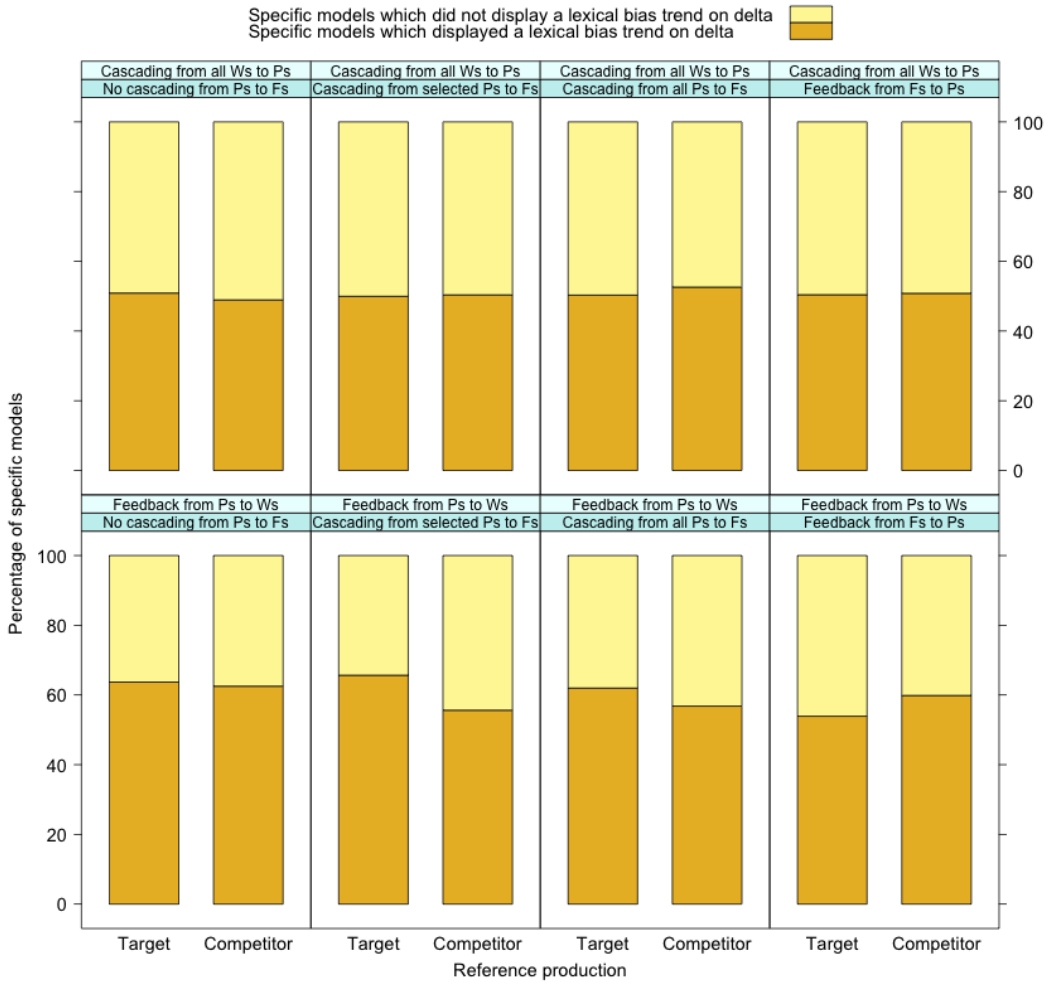


Figure 8.5: The effect of modifying activation flow on whether two-stage models display a numerically smaller delta for stimuli with a lexical error outcome, for vectors of alveolar and velar feature activation values used to simulate tongue-to-palate contact.
Key: Ws = words, Ps = phonemes, Fs = features

specific models can truly account for this effect. Furthermore, with the current numerical trend statistics where the probability of finding a trend by chance is 0.5, the power of a binomial analysis attempting to find evidence of multiple effects would be extremely low.

The effect of spreading activation parameter manipulations on lexical bias trends

In the same way that we have previously analysed the effect of spreading activation parameter manipulations on whether various effects were significant or not, we can also investigate the effect of parameter manipulations on the number of models for which there is a numerically smaller average delta in the lexical outcome condition,

Table 8.3: Binomial analysis to determine which two-stage architectures have a significant number of models displaying a numerically smaller delta measured from the competitor reference for stimuli with a lexical error outcome, for vectors of alveolar and velar feature activation values used to simulate tongue-to-palate contact. An asterisk indicates that there would be less than 0.025 probability (Bonferroni corrected to 0.003125) of witnessing the same number or more specific models generating lexical bias trends by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	LB trend on delta		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	2916	1427	0.871	
Cascading from selected Ps to Fs	2916	2916	1469	0.335	
Cascading from all Ps to Fs	2916	2916	1534	0.002	*
Feedback from Fs to Ps	5832	5832	2964	0.102	
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	5832	3646	$< .001$	*
Cascading from selected Ps to Fs	5832	5832	3244	$< .001$	*
Cascading from all Ps to Fs	5832	5832	3312	$< .001$	*
Feedback from Fs to Ps	5832	5832	3491	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

LB = lexical bias, Prob. = probability

Table 8.4: Binomial analysis to determine which two-stage architectures have a significant number of models displaying a numerically smaller delta measured from the target reference for stimuli with a lexical error outcome, for vectors of alveolar and velar feature activation values used to simulate tongue-to-palate contact. An asterisk indicates that there would be less than 0.025 probability (Bonferroni corrected to 0.003125) of witnessing the same number or more specific models generating lexical bias trends by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	LB trend on delta		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	2916	1484	0.163	
Cascading from selected Ps to Fs	2916	2916	1457	0.507	
Cascading from all Ps to Fs	2916	2916	1467	0.362	
Feedback from Fs to Ps	5832	5832	2941	0.252	
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	5832	3714	$< .001$	*
Cascading from selected Ps to Fs	5832	5832	3830	$< .001$	*
Cascading from all Ps to Fs	5832	5832	3615	$< .001$	*
Feedback from Fs to Ps	5832	5832	3143	$< .001$	*

Key:

Ws = words, Ps = phonemes, Fs = features

LB = lexical bias, Prob. = probability

as shown in table 8.5 and figure 8.6. As we have previously argued that this effect can largely be explained by more errors at the phoneme level occurring in the lexical outcome condition, it makes sense that a number of parameters which we know increase error rate also increase the number of models which trend towards a lexical bias effect: namely, low jolt to prime ratio, high decay and high activation-based noise levels. We also find more models which trend towards a lexical bias effect when there are higher numbers of steps before selection. This is in line with the theoretical observation that a lexical bias effect on phoneme selection due to feedback would be impossible with fewer than three steps before selection. However, despite further results in section 7.3 showing that lexical bias increases as the number of steps before selection increase, figure 8.6 does not give any clear indication that having eight steps before selection is more conducive to this effect than having five steps before selection. On closer examination, it is not clear that the highest activation-based noise level offers any advantage over the medium activation-based noise level, or that the lowest jolt to prime ratio is more conducive to this effect than the second lowest jolt to prime ratio.

Similarly, it is not clear that extreme values of connection strength are optimal for a lexical bias effect to be detected on delta measurements. Table 8.5 somewhat surprisingly suggests that lower connection strengths lead to more models demonstrating an effect in the desired direction, despite the fact that we showed in chapter 7 that higher connectivity strengths boost lexical bias. Figure 8.6 provides more information and suggests that there is in fact a happy medium, such that increasing forward and feedback connection strength first supports the lexical bias on delta effect, but then acts against it as connection strength becomes very strong. This may be because the frequency effect grows faster with relation to connection strength than the lexical bias effect does, as the frequency effect is driven by many word nodes rather than a single node like the lexical bias effect. Equally, very high numbers of steps before selection, very high levels of activation-based noise, and very low jolt to prime ratios may offer too much support to the activation flow driven frequency effect. We note that suppression of activation flow to maintain a happy medium may be an additional reason why high decay rates are advantageous when simulating this result.

Finally, we note that there are a few more models with effects in the right direction when levels of intrinsic noise are low. One possible explanation of this result may be that intrinsic noise does not affect activation levels enough to cause many more

Table 8.5: Results of logistic regression model analyses using parameter values to predict whether two-stage models with feedback from phonemes to words display a numerically smaller delta measured from the competitor reference for stimuli with a lexical error outcome, for vectors of alveolar and velar feature activation values used to simulate tongue-to-palate contact. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	−	4.2	18	< .001	*
joltPrimeRatio	−	11.0	121	< .001	*
decay	+	5.4	30	< .001	*
steps	+	18.7	353	< .001	*
actiNoiseSD	+	6.7	45	< .001	*
intrinNoiseSD	−	2.2	5	0.026	*

errors, but it does slightly distort the activation patterns transmitted at phoneme selection before feature selection occurs.

8.2.3 Conclusions

Our investigations have found no evidence that any specific model of any architecture can demonstrate a significantly smaller delta from the reference measurement for the competing place of articulation in the lexical outcome condition in comparison to the non-lexical outcome condition, as found by McMillan et al. (2009). This appears to be largely due to extremely strong effects of onset and coda frequency which outweigh the effects of lexicality. However, we did find that architectures with feedback from phonemes to words demonstrated a significant number of models showing lexical bias trends on delta measurements, regardless of activation flow between phonemes and features. This fits in with our predictions that no cascading from phonemes would be required to account for this result. Finally, we showed that parameters which increase error rate at the phoneme level support this effect. This is also true for parameters which increase interactivity, to an extent. However, too much interactivity appears to reduce the number of models demonstrating the lexical bias effect on delta, probably because this makes the frequency effect far too strong for the lexicality effect to be detected.

Our first attempts to model EPG within Dell's (1986) architecture have therefore had limited success, as frequency has exhibited such a strong effect on the activation of both the alveolar and velar features. Our VOT measure was perhaps not affected quite so strongly by this variable because it involves subtracting one feature

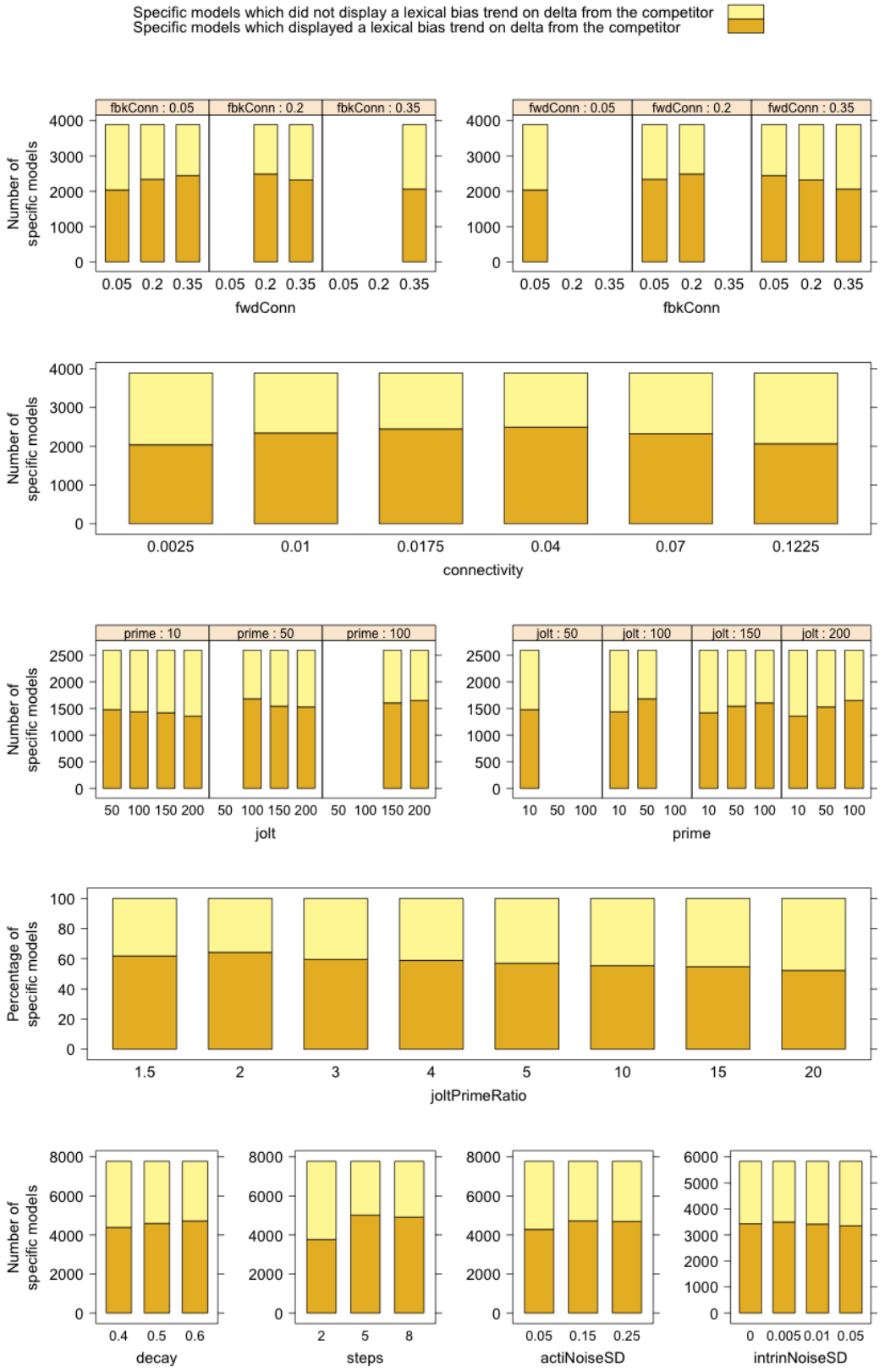


Figure 8.6: The effect of parameter manipulations on whether two-stage models with feedback from phonemes to words display a numerically smaller delta measured from the competitor reference for stimuli with a lexical error outcome, for vectors of alveolar and velar feature activation values used to simulate tongue-to-palate contact.

activation value from another, therefore reducing the effects of any general increase in activation affecting both features.

8.3 McMillan's (2008) evidence of a phonological similarity effect on articulatory and acoustic measurements

In this final section, we investigate whether any of the models can account for McMillan's (2008) result that VOT measurements of articulations are further from a reference VOT measurement for the target onset when the competing onset differs only in its voicing feature, rather than both its place and voicing feature; and similarly, whether articulations measured by EPG and ultrasound are further from a reference EPG/ultrasound measurement for the target onset when the competing onset shares differs only in its place feature, rather than both its place and voicing feature. We expect to find support for McMillan's (2008) claim that feedback from features to phonemes is required to account for this result. In chapter 7 we showed that feedback from features to phonemes is not required to account for the transcribed phonological similarity effect when output is at the featural level, and that even a model with no cascading from phonemes can account for this result when enough featural level contextual errors occur. We further demonstrated that Goldrick and Blumstein's (2006) findings and McMillan et al.'s (2009) results do not place stronger constraints on activation flow between features and phonemes. If these simulations provide support for the hypothesis that feedback from features to phonemes is required to account for McMillan's (2008) findings, this result would constitute the sole constraint on phoneme-to-feature activation flow. Our discovery in the previous section that frequency exerts such a strong effect on the EPG simulations does not bode well for our EPG/ultrasound results, however.

8.3.1 Simulation methodology

In this first investigation of McMillan's (2008) delta evidence for the effects of phonological similarity, we carried out a delta analysis of the output of the phonological similarity simulations described in section 7.3.

Model configuration

All 37,908 two-stage models were tested.

Model task and lexicon

As described in section 7.3, the model produced single words while competitor words were primed. The target materials and 100 word lexicon are described in full in section 6.2.2. Reference utterances were also generated, as described in section 8.2.1. In this study, we consider behaviour from simulations using the material set in which the place of articulation of the target and competitor onset always differs (to allow us to simulate McMillan's (2008) EPG and ultrasound analyses), and behaviour from simulations using the material set in which voicing of the target and competitor onset always differs (to allow us to simulate McMillan's (2008) VOT analysis).

Model output interpretation

To analyse output for comparison to the EPG and ultrasound results reported by McMillan (2008), we again recorded the activation of the alveolar and velar features, which we interpreted to represent the extent to which the resulting articulation involves tongue raising at the front and the back of the mouth. We assume that this is reflected by tongue height in ultrasound measurements.

Output from the productions in which a target word received both jolt and prime activation was used to calculate reference alveolar and velar activation vectors for each onset in the material set. For each production of the target words with primed competitors, the Euclidean distance between the alveolar and velar feature activation vector recorded from that production and the reference activation vector for the target onset was calculated to represent delta. A t-test comparison of this delta measurement in the condition where target and competitor onsets differed in place only, and the condition where target and competitor onsets differed in both place and voicing, allowed us to determine which specific models displayed a bigger delta when onset consonants were more similar, as reported by McMillan (2008).

To analyse output for comparison to McMillan's (2008) VOT results, we calculated a value to simulate VOT by subtracting the activation of the voiced feature from the voiceless feature, as in the previous chapter.

Output from the productions in which a target word received both jolt and prime activation was used to calculate reference VOTs for each onset in the material set. For each production of the target words with primed competitors, the absolute difference between the VOT recorded from that production and the reference VOT

for the target onset was calculated to represent delta. A t-test comparison of this delta measurement in the condition where target and competitor onsets differed in voicing only, and the condition where target and competitor onsets differed in both place and voicing, allowed us to determine which specific models displayed a bigger delta when onset consonants were more similar, as reported by McMillan (2008).

8.3.2 Simulation results

We begin our analysis by investigating which architectures can account for McMillan's (2008) results.

Architecture analysis of phonological similarity effects

We first consider results for our VOT simulations. Table 8.6 shows that the architecture with no feedback from phonemes to words and feedback from features to phonemes shows significantly smaller delta measurements in the condition where the competing onset differs in voicing feature only, in comparison to the condition where the competing onset differs in both voicing and place feature. However, contrary to our predictions, no evidence is found that the architecture with feedback from phonemes to words and feedback from features to phonemes can account for McMillan's (2008) findings.

However, figure 8.7 shows that very few specific models with no feedback from phonemes to words and feedback from features to phonemes show significant effects. This again suggests that the effect is possibly weak. The effect would probably be most visible in models in which high numbers of contextual errors at the phoneme level occur, as these errors would help support the effect by exaggerating the activation differences between productions in the phonologically similar condition and productions in the phonologically dissimilar condition. Correspondingly, table 8.7 and figure 8.8 show that when we exclude models which generate too many errors or too high a proportion of non-contextual errors for the constraints determined in chapter 4, there is no evidence that any architecture can capture this effect.

As in the previous section, we find that no significant results are found for any architecture in our EPG/ultrasound simulations, as shown in table 8.8 and figure 8.7. In the next section, we look for possible effects of frequency in the architectures with feedback from features to phonemes to determine whether these are again responsible for the lack of significant results.

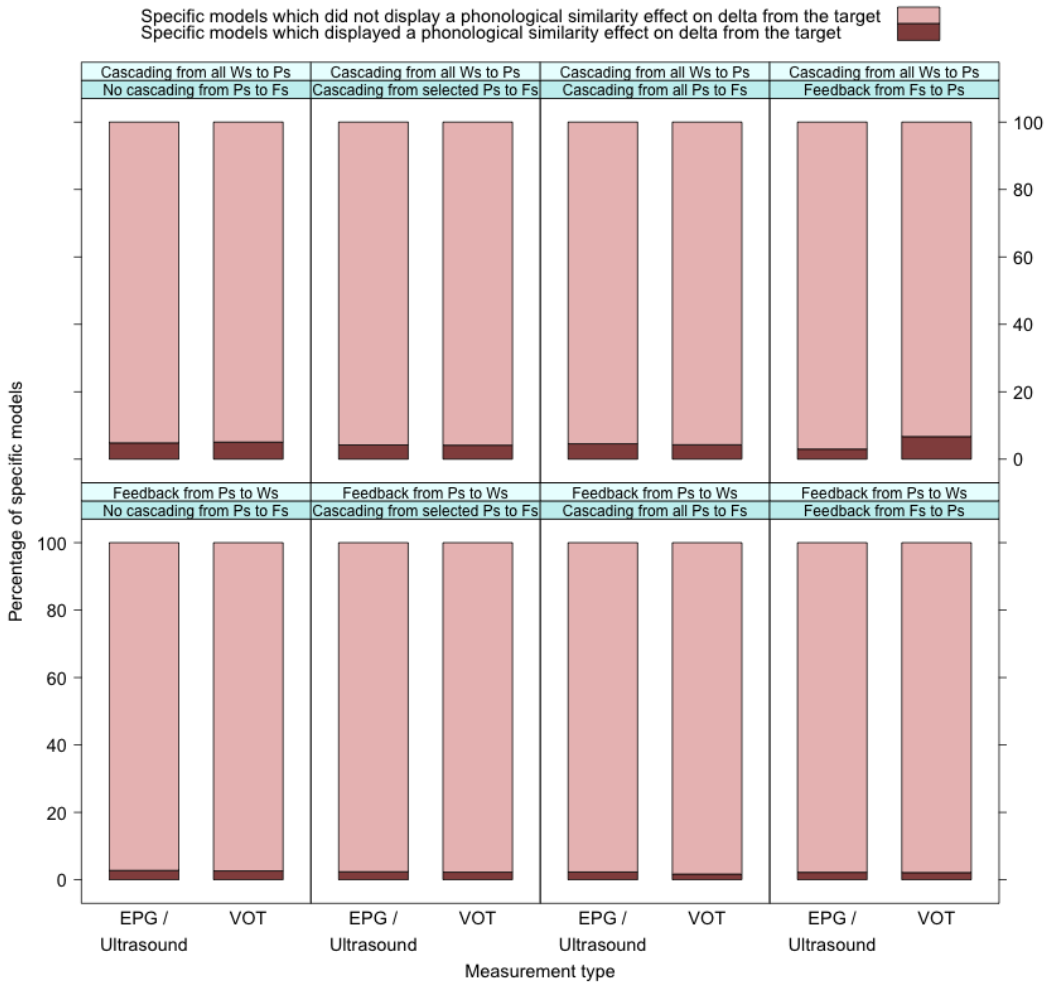


Figure 8.7: The effect of modifying activation flow on whether two-stage models display a larger delta measured from the target reference for stimuli where the competing onset was similar, for simulated VOT values and vectors of alveolar and velar feature activation values used to simulate tongue-to-palate contact or tongue height.
Key: Ws = words, Ps = phonemes, Fs = features.

Table 8.6: Binomial analysis to determine which two-stage architectures display a larger delta measured from the target reference for stimuli where the competing onset was similar, for simulated VOT values. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating phonological similarity effects by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	PS effect on delta		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	2916	148	0.405	
Cascading from selected Ps to Fs	2916	2916	121	$> .9$	
Cascading from all Ps to Fs	2916	2916	125	$> .9$	
Feedback from Fs to Ps	5832	5832	391	$< .001$	*
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	5832	153	$> .9$	
Cascading from selected Ps to Fs	5832	5832	131	$> .9$	
Cascading from all Ps to Fs	5832	5832	99	$> .9$	
Feedback from Fs to Ps	5832	5832	125	$> .9$	

Key:

Ws = words, Ps = phonemes, Fs = features

PS = phonological similarity, Prob. = probability

Table 8.7: Binomial analysis to determine which two-stage architectures display a larger delta measured from the target reference for stimuli where the competing onset was similar, for simulated VOT values, excluding specific models that do not pass both constraints on error rate and non-contextuality of errors. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating phonological similarity effects by chance.

	Specific model counts				**Prob.**
	Total	Excluded	Sufficient data	PS effect on delta	
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	802	2114	115	0.164
Cascading from selected Ps to Fs	2916	819	2097	77	$> .9$
Cascading from all Ps to Fs	2916	830	2086	87	$> .9$
Feedback from Fs to Ps	5832	2996	2836	107	$> .9$
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	2503	3329	124	$> .9$
Cascading from selected Ps to Fs	5832	2488	3344	101	$> .9$
Cascading from all Ps to Fs	5832	2827	3005	70	$> .9$
Feedback from Fs to Ps	5832	3199	2633	82	$> .9$

Key:

Ws = words, Ps = phonemes, Fs = features

PS = phonological similarity, Prob. = probability

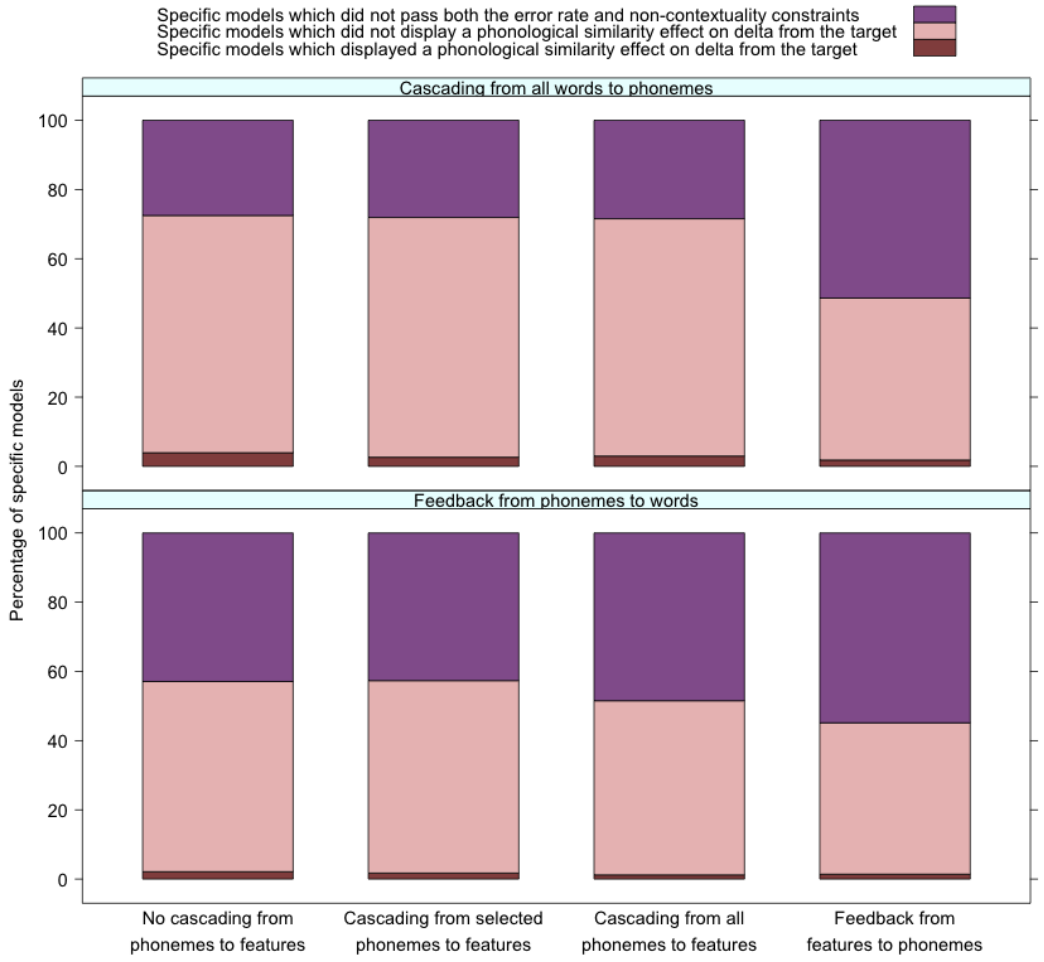


Figure 8.8: The effect of modifying activation flow on whether two-stage models display a larger delta measured from the target reference for stimuli where the competing onset was similar, for simulated VOT values, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

Table 8.8: Binomial analysis to determine which two-stage architectures display a larger delta measured from the target reference for stimuli where the competing onset was similar, for vectors of alveolar and velar feature activation values used to simulate tongue-to-palate contact or tongue height. An asterisk indicates that there would be less than 0.05 probability (Bonferroni corrected to 0.00625) of witnessing the same number or more specific models generating phonological similarity effects by chance.

	Specific model counts			**Prob.**
	Total	Sufficient data	PS effect on delta	
Cascading from all Ws to Ps				
No cascading from Ps to Fs	2916	2916	141	0.638
Cascading from selected Ps to Fs	2916	2916	123	$> .9$
Cascading from all Ps to Fs	2916	2916	133	0.852
Feedback from Fs to Ps	5832	5832	175	$> .9$
Feedback from Ps to Ws				
No cascading from Ps to Fs	5832	5832	162	$> .9$
Cascading from selected Ps to Fs	5832	5832	139	$> .9$
Cascading from all Ps to Fs	5832	5832	134	$> .9$
Feedback from Fs to Ps	5832	5832	129	$> .9$

Key:

Ws = words, Ps = phonemes, Fs = features

PS = phonological similarity, Prob. = probability

Table 8.9: Frequency of occurrence of stimulus place-voicing onset feature combinations in the model's phoneme inventory.

Place-voicing combination	**Frequency**
alveolar voiced	5
alveolar voiceless	2
velar voiced	1
velar voiceless	1

Place and voicing feature activation

Figure 8.9 shows average simulated VOT values for the architecture with no feedback from phonemes to words, and feedback from features to phonemes, for which we found a significant effect of phonological similarity. Even here effects of frequency are visible, although this time these are feature frequency effects rather than phoneme frequency effects, as there is feedback from features to phonemes but no feedback from phonemes to words. Most noticeably, the voiced feature is clearly more activated than the voiceless feature in nearly all of the average productions, as their simulated VOTs are below 0. This reflects the fact that there are 12 voiced onset phonemes in the lexicon, in comparison to 9 voiceless onset phonemes.

It can also be seen however that alveolar onsets are generally more voiced. There are 7 alveolar onset phonemes in the lexicon in comparison to 2 velar onset phonemes. Alveolar onsets will generally become more activated therefore, and this activation will be multiplied more by feedback loops to the voiced feature than by the voiceless feature, due to the greater frequency of the voiced feature. In addition, the alveolar and voiced features occur together more frequently than other relevant place and voicing feature combinations, as can be seen in table 8.9, so activation of one feature will support activation of the other. The lower VOT of alveolar onsets is very visible on target onsets, but can be seen on competitor onsets too.

Figure 8.10 shows that similar results are found for alveolar and velar feature activation values in the architecture with no feedback from phonemes to words, and feedback from features to phonemes. There are more onset phonemes with a voiced feature than a voiceless feature, and so voiced onsets display more alveolar and velar activation. Again, the effect is most obvious on target onsets but can also be observed on competitor onsets. Finally, there is a greater effect of the voiced onset activation boost on alveolar activation, because there are more alveolar onsets than velar onsets, and more alveolar voiced onsets than velar voiced onsets.

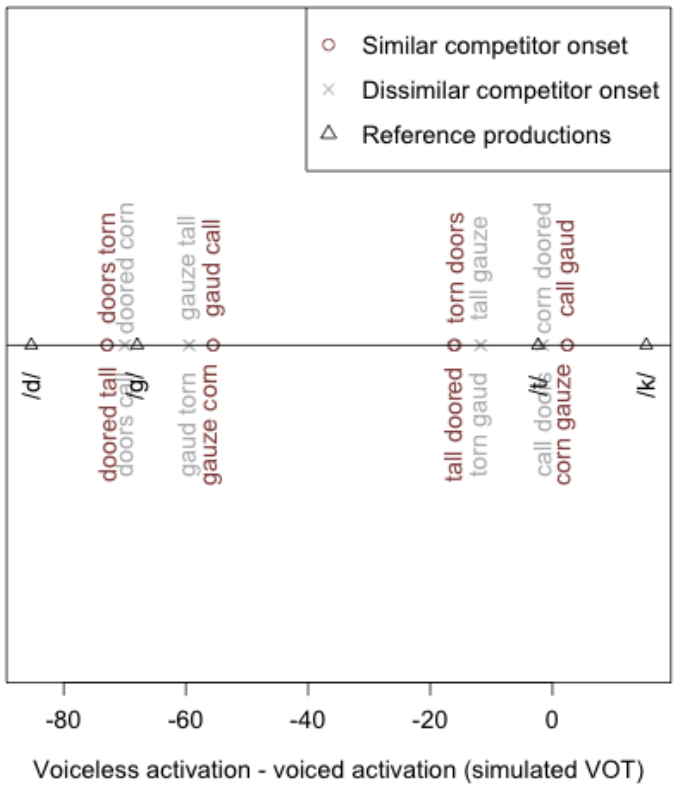


Figure 8.9: Mean simulated VOT values for productions of single words when competitor words are primed, in two-stage models with no feedback from phonemes to words and feedback from features to phonemes.

Once feedback from phonemes to words is added to the feedback from features to phonemes, effects of phoneme frequency can be observed again, as evident in the diagram of VOT values in figure 8.11. The activation boost caused by the presence of the very frequent /n/ coda is particularly noticeable, and causes further activation of the frequent voiced onset feature when found in both target and competitor words. The effect of this coda can also be observed on alveolar and velar feature activation values in figure 8.12. In this figure, there are also hints of the effect of onset frequency. Specifically, /k/ is the most frequent onset, although its voiceless feature is less frequent than the voiced feature. The diagram gives some indication that voiceless productions tend to be more /k/-like (i.e., have more velar activation) than their voiced counterparts.

It is possible therefore that the extra variance due to phoneme frequency is preventing a significant phonological similarity result on VOT results for the architecture with feedback from phonemes to words, and feedback from features to

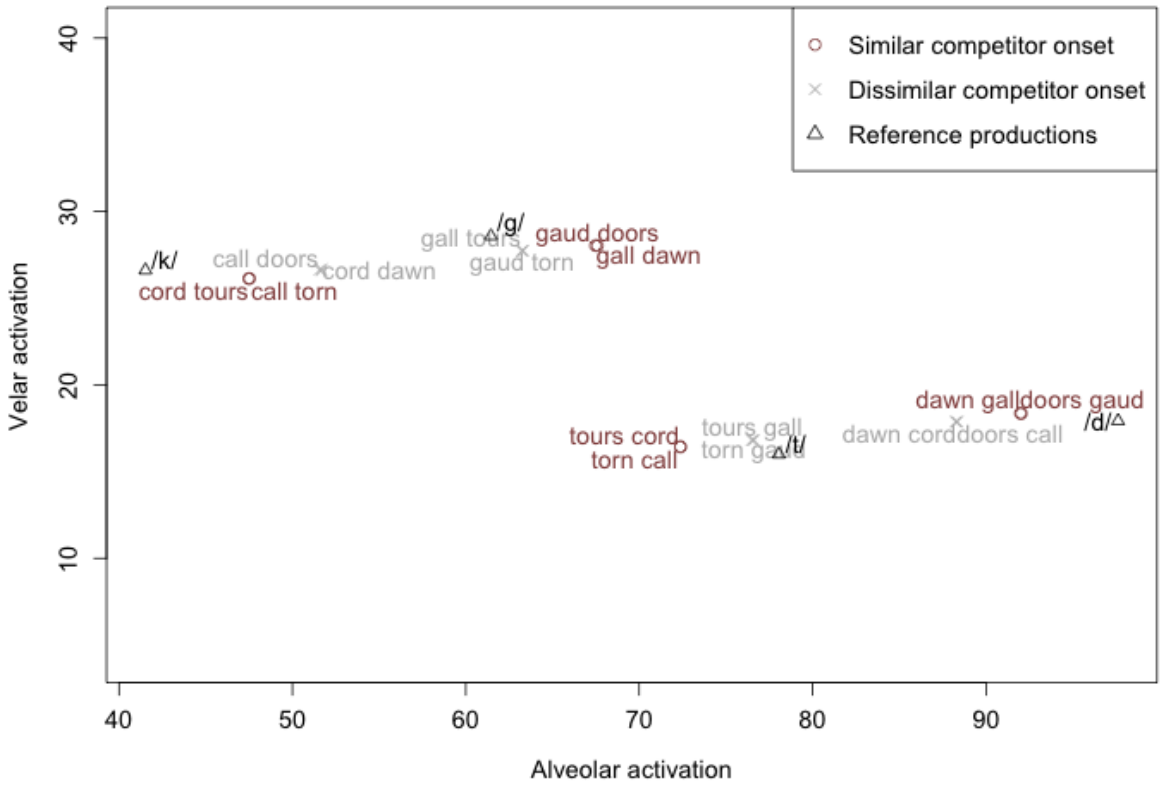


Figure 8.10: Mean alveolar and velar activation values as used to simulate tongue height, for productions of single words when competitor words are primed, in two-stage models with no feedback from phonemes to words and feedback from features to phonemes. The same distance is used to represent one unit of activation on the x-axis as on the y-axis.

phonemes. Again, we suggest that frequency causes more of a problem overall in the EPG/ultrasound simulations, because the subtraction calculation in the VOT measure reduces the effect of the general activation increase caused by more frequent phonemes. There is still a possibility that architectures with feedback from features to phonemes are exhibiting results trending in the right direction on both measures however, and in the next section we investigate whether we can find evidence for such trends.

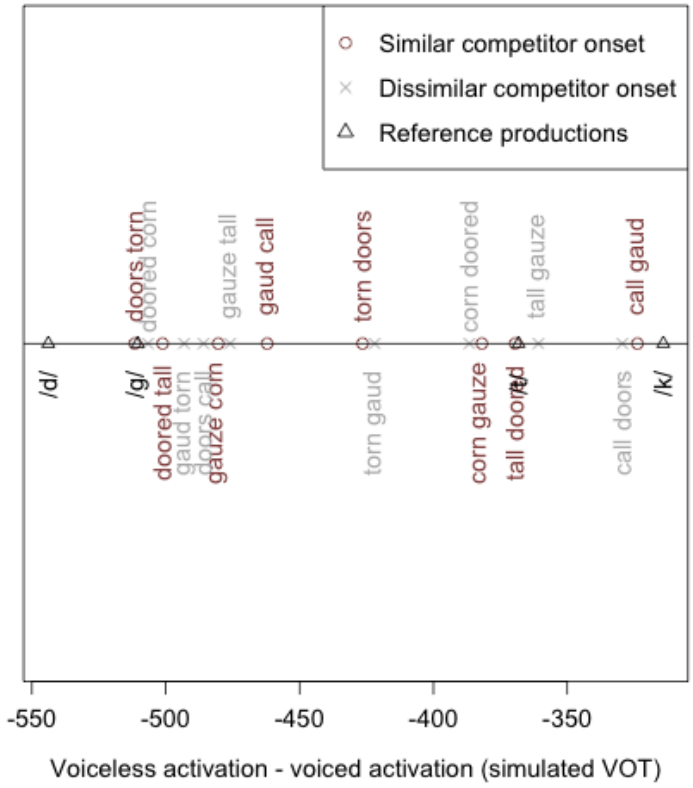


Figure 8.11: Mean simulated VOT values for productions of single words when competitor words are primed, in two-stage models with feedback from phonemes to words and from features to phonemes.

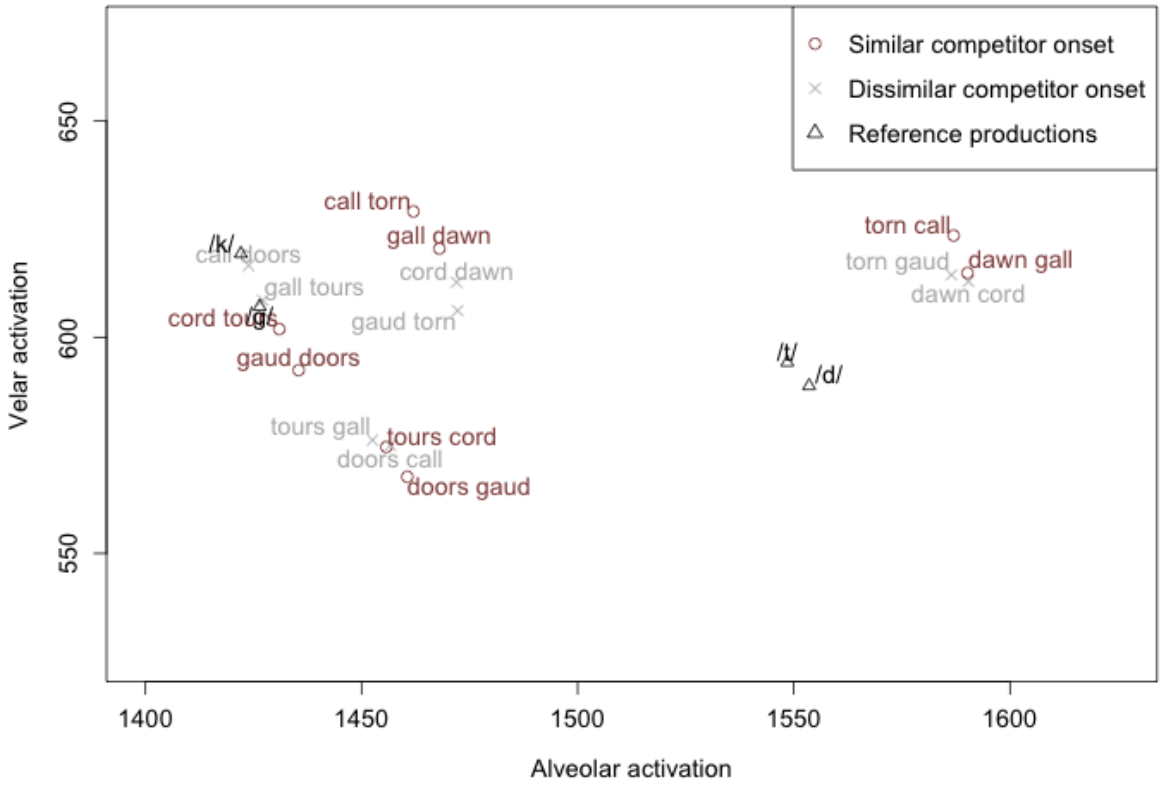


Figure 8.12: Mean alveolar and velar activation values as used to simulate tongue height, for productions of single words when competitor words are primed, in two-stage models with feedback from phonemes to words and feedback from features to phonemes. The same distance is used to represent one unit of activation on the x-axis as on the y-axis.

Table 8.10: Binomial analysis to determine which two-stage architectures have a significant number of models displaying a numerically larger delta measured from the target reference for stimuli where the competing onset was similar, for simulated VOT values. An asterisk indicates that there would be less than 0.025 probability (Bonferroni corrected to 0.003125) of witnessing the same number or more specific models generating phonological similarity trends by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	PS trend on delta		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	2916	1471	0.309	
Cascading from selected Ps to Fs	2916	2916	1424	0.893	
Cascading from all Ps to Fs	2916	2916	1456	0.522	
Feedback from Fs to Ps	5832	5832	3573	$< .001$	*
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	5832	2252	$> .9$	
Cascading from selected Ps to Fs	5832	5832	2248	$> .9$	
Cascading from all Ps to Fs	5832	5832	2430	$> .9$	
Feedback from Fs to Ps	5832	5832	2873	0.867	

Key:

Ws = words, Ps = phonemes, Fs = features

PS = phonological similarity, Prob. = probability

Architecture analysis of phonological similarity trends

Table 8.10 shows trend analysis results for the VOT measurements. Within the architecture with no feedback from phonemes to words and feedback from features to phonemes, there are more specific models than would be predicted by chance displaying a numerically smaller delta from the reference target onset in the condition where the competitor onset differs only in voicing feature, in comparison to the condition where the competitor onset differs in both voicing and place feature. However, there is again no evidence for such a trend in models with feedback from phonemes to words. Table 8.11 shows the EPG/ultrasound measurements. Similarly, there are more specific models than would be predicted by chance displaying a numerically smaller delta from the reference target onset in the condition where the competitor onset differs only in place feature, in comparison to the condition where the competitor onset differs in both voicing and place feature, for the architecture with no feedback from phonemes to words and feedback from features to phonemes. Again however, no evidence is found for this trend in models with feedback from phonemes to words.

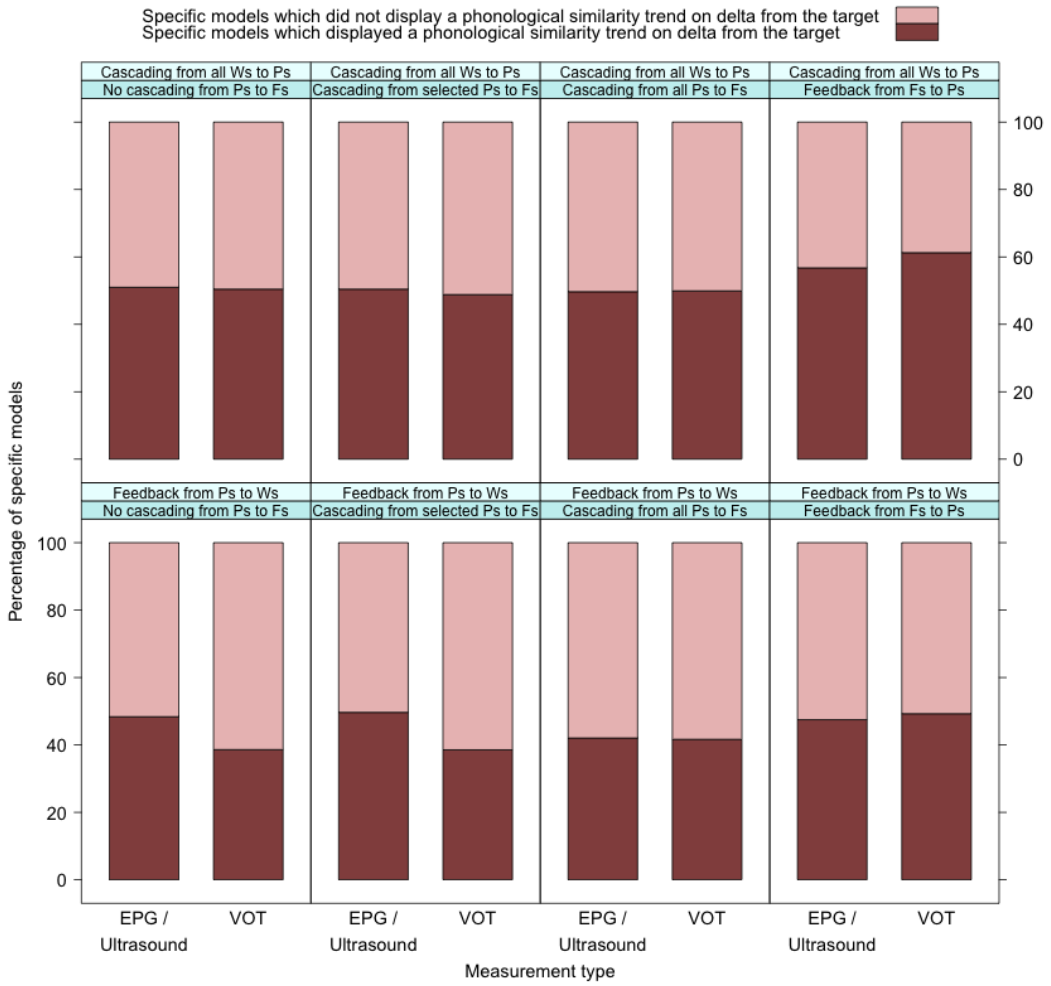


Figure 8.13: The effect of modifying activation flow on whether two-stage models display a larger delta measured from the target reference for stimuli where the competing onset was similar, for simulated VOT values and vectors of alveolar and velar feature activation values used to simulate tongue-to-palate contact or tongue height.
Key: Ws = words, Ps = phonemes, Fs = features

Table 8.11: Binomial analysis to determine which two-stage architectures have a significant number of models displaying a numerically larger delta measured from the target reference for stimuli where the competing onset was similar, for vectors of alveolar and velar feature activation values used to simulate tongue-to-palate contact or tongue height. An asterisk indicates that there would be less than 0.025 probability (Bonferroni corrected to 0.003125) of witnessing the same number or more specific models generating phonological similarity trends by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	PS trend on delta		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	2916	1487	0.137	
Cascading from selected Ps to Fs	2916	2916	1471	0.309	
Cascading from all Ps to Fs	2916	2916	1449	0.624	
Feedback from Fs to Ps	5832	5832	3310	$< .001$	*
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	5832	2821	$> .9$	
Cascading from selected Ps to Fs	5832	5832	2896	0.695	
Cascading from all Ps to Fs	5832	5832	2453	$> .9$	
Feedback from Fs to Ps	5832	5832	2769	$> .9$	

Key:
Ws = words, Ps = phonemes, Fs = features
PS = phonological similarity, Prob. = probability

Furthermore, figure 8.13 very surprisingly suggests that in some architectures with feedback from phonemes to words, there may be a reverse phonological similarity effect, for both measures. The analysis presented in table 8.12 confirms that for architectures with feedback from phonemes to words and without feedback from features to phonemes, there is a reverse phonological similarity effect on VOT measurements analysed using the delta measure. Table 8.13 shows that for EPG/ultrasound measurements analysed using the delta method, there is a reverse phonological similarity effect for architectures with feedback from phonemes to words and cascading from all phonemes to features, or architectures with feedback from features to phonemes.

It is not at all clear why feedback from phonemes to words should have this effect. The only reasonable suggestion apparent at this point would be that there is a confound in the material set design, which was described in section 6.2.2. There are no obvious candidates for such a confound, given all the variables controlled for in the design of this material set. Perhaps the first possibility for investigation would be the fact that the biggest difference in onset phoneme frequency exists between the onsets /k/ (which occurs 6 times in the lexicon) and /d/ (which occurs 4 times in the lexicon), and these two phonemes differ in both place and voicing. It

Table 8.12: Binomial analysis to determine which two-stage architectures have a significant number of models displaying a numerically smaller delta measured from the target reference for stimuli where the competing onset was similar, for simulated VOT values. An asterisk indicates that there would be less than 0.025 probability (Bonferroni corrected to 0.003125) of witnessing the same number or more specific models generating reverse phonological similarity trends by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	Reverse PS trend on delta		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	2916	1445	0.678	
Cascading from selected Ps to Fs	2916	2916	1492	0.101	
Cascading from all Ps to Fs	2916	2916	1460	0.463	
Feedback from Fs to Ps	5832	5832	2259	$> .9$	
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	5832	3580	$< .001$	*
Cascading from selected Ps to Fs	5832	5832	3584	$< .001$	*
Cascading from all Ps to Fs	5832	5832	3402	$< .001$	*
Feedback from Fs to Ps	5832	5832	2959	0.127	

Key:
Ws = words, Ps = phonemes, Fs = features
PS = phonological similarity, Prob. = probability

was assumed that any effects of difference in frequency would balance out through use of all phonemes as target and competitors. For example, whilst a target with onset /d/ and competitor /k/ is likely to display large deltas from the target due to the activation of competitor /k/ and its features becoming high in comparison to the activation of target /d/ and its features, it would seem logical that a target with onset /k/ and competitor /d/ would compensate with low deltas from the target due to the activation of competitor /k/ and its features becoming high in comparison to the activation of target /d/ and its features. Further investigations could establish whether this is in fact not the case.

If such a confound did exist, this would further explain why no phonological similarity effect in the expected direction is found in the architecture with feedback from phonemes to words and feedback from features to phonemes.

We suggest that an analysis of which parameter settings allow the implementation to capture both this effect and others which we have modelled would be more useful once the problems caused by frequency have been addressed.

Table 8.13: Binomial analysis to determine which two-stage architectures have a significant number of models displaying a numerically smaller delta measured from the target reference for stimuli where the competing onset was similar, for vectors of alveolar and velar feature activation values used to simulate tongue-to-palate contact or tongue height. An asterisk indicates that there would be less than 0.025 probability (Bonferroni corrected to 0.003125) of witnessing the same number or more specific models generating reverse phonological similarity trends by chance.

	Specific model counts			**Prob.**	
	Total	Sufficient data	Reverse PS trend on delta		
Cascading from all Ws to Ps					
No cascading from Ps to Fs	2916	2916	1429	0.854	
Cascading from selected Ps to Fs	2916	2916	1445	0.678	
Cascading from all Ps to Fs	2916	2916	1467	0.362	
Feedback from Fs to Ps	5832	5832	2522	$> .9$	
Feedback from Ps to Ws					
No cascading from Ps to Fs	5832	5832	3011	0.006	
Cascading from selected Ps to Fs	5832	5832	2936	0.296	
Cascading from all Ps to Fs	5832	5832	3379	$< .001$	*
Feedback from Fs to Ps	5832	5832	3063	$< .001$	*

Key:
Ws = words, Ps = phonemes, Fs = features
PS = phonological similarity, Prob. = probability

The effect of spreading activation parameter manipulations on phonological similarity effects

Despite these concerns about possible confounds in the architecture with feedback from phonemes to words, we will investigate how parameter manipulations affect whether the architecture with no feedback from phonemes to words and feedback from features to phonemes displays a phonological similarity effect on delta. Findings are similar regardless of whether we consider EPG/ultrasound or VOT output, so here we present an investigation of the effect of parameter manipulations on significant phonological similarity effects for VOT measurements.

Our results are very similar to those found for the lexical bias effect on delta. Table 8.14 shows that specific models with low jolt to prime ratios and high levels of activation-based noise are more likely to display phonological similarity effects on delta, potentially because these parameters lead to higher error rates. However, figure 8.14 indicates that a low jolt to prime ratio of 2 may be superior to a very low jolt to prime ratio of 1.5, and that a high activation-based noise level of 0.15 may be superior to a very high activation-based noise level of 0.25.

An increase in steps before selection also boosts the number of models showing a significant effect. Whilst a phonological similarity effect would be impossible on phoneme selection given less than three steps before selection, it is unclear from figure 8.14 that there is any further advantage of higher numbers of steps before selection.

Finally, whilst table 8.14 suggests that connection strengths have no influence on phonological similarity effects, figure 8.14 shows that there is once again a happy medium. Very low connection strengths do not lead to many models exhibiting effects in the right direction, medium connection strengths display higher numbers of models showing effects, but at very high connection strengths, the number of models exhibiting effects begins to reduce again.

As with the lexical bias effect on delta, these results which demonstrate a need for moderation in parameter settings may well be due to the frequency effect growing faster than the phonological similarity effect as activation flow through the network increases. This is because frequency effects are driven by many phoneme nodes, whereas the phonological similarity effect relies on one feature node only.

For these results, the effect of decay is only marginal, but it is in the same direction as for the lexical bias effect on delta, such that higher levels of decay lead to more models displaying phonological similarity effects. This may be due either to error rates rising at higher decay rates, or because of the reduction of activation flow which higher decay rates would cause in turn helps reduce the influence of frequency effects. There is no significant effect of intrinsic noise.

Figure 8.15 shows that when the constraints on error rate and non-contextuality of errors are applied, models with high connection strengths, a low jolt to prime ratio, a high level activation-based noise, and a high number of steps before selection, are ruled out. Removal of these high error rate specific models means that very few models which display a phonological similarity effect remain.

8.3.3 Conclusions

McMillan's (2008) finding constitutes the first result which we have simulated which cannot be accounted for by all phoneme-to-feature activation flow options. It was shown that an architecture with feedback from features to phonemes and no feedback from phonemes to words can account for this result on VOT measurements, but no evidence was found to demonstrate that an architecture with feedback from

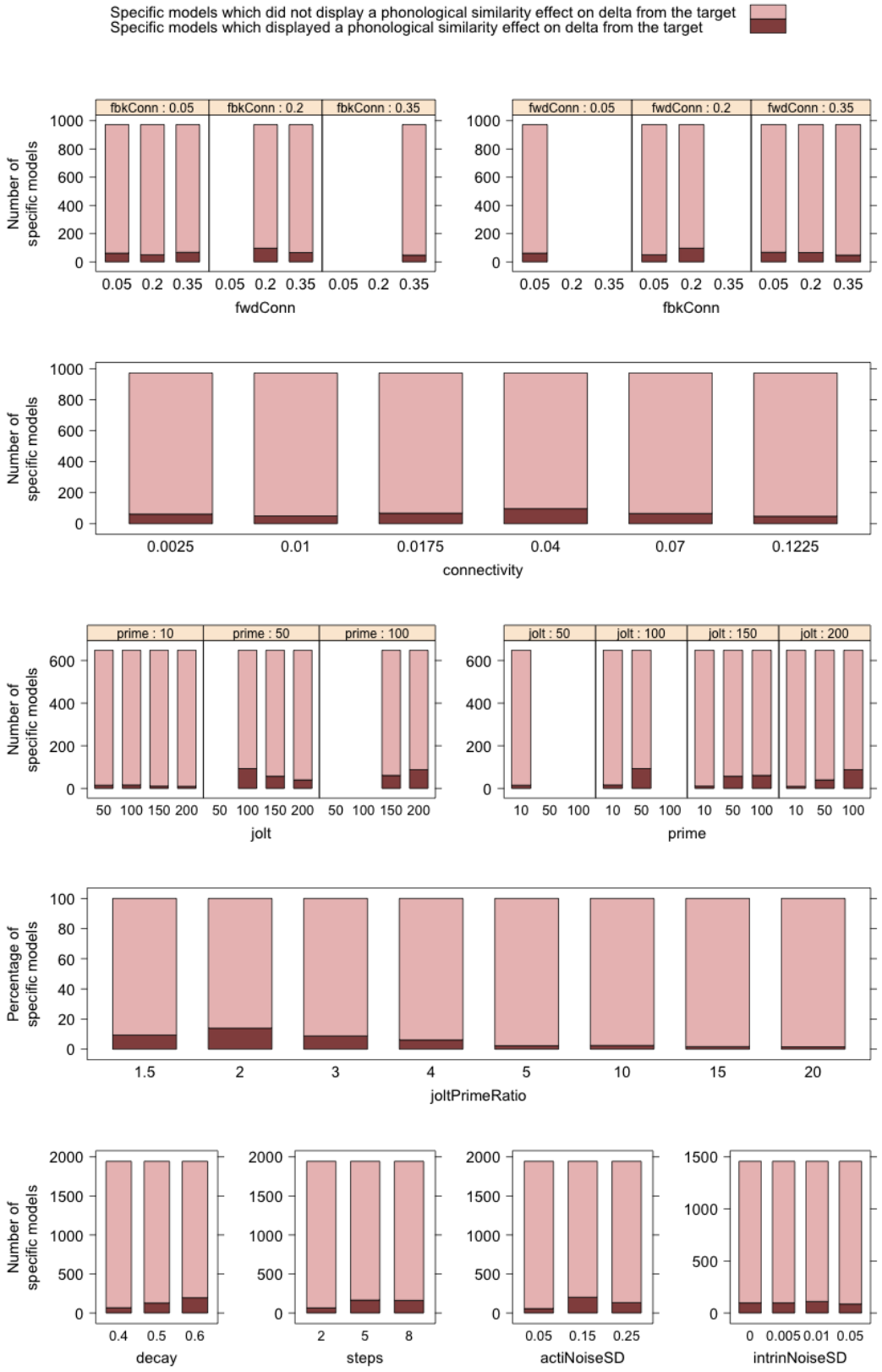


Figure 8.14: The effect of parameter manipulations on whether two-stage models with no feedback from phonemes to words and feedback from features to phonemes display a larger delta measured from the target reference for stimuli where the competing onset was similar, for simulated VOT values.

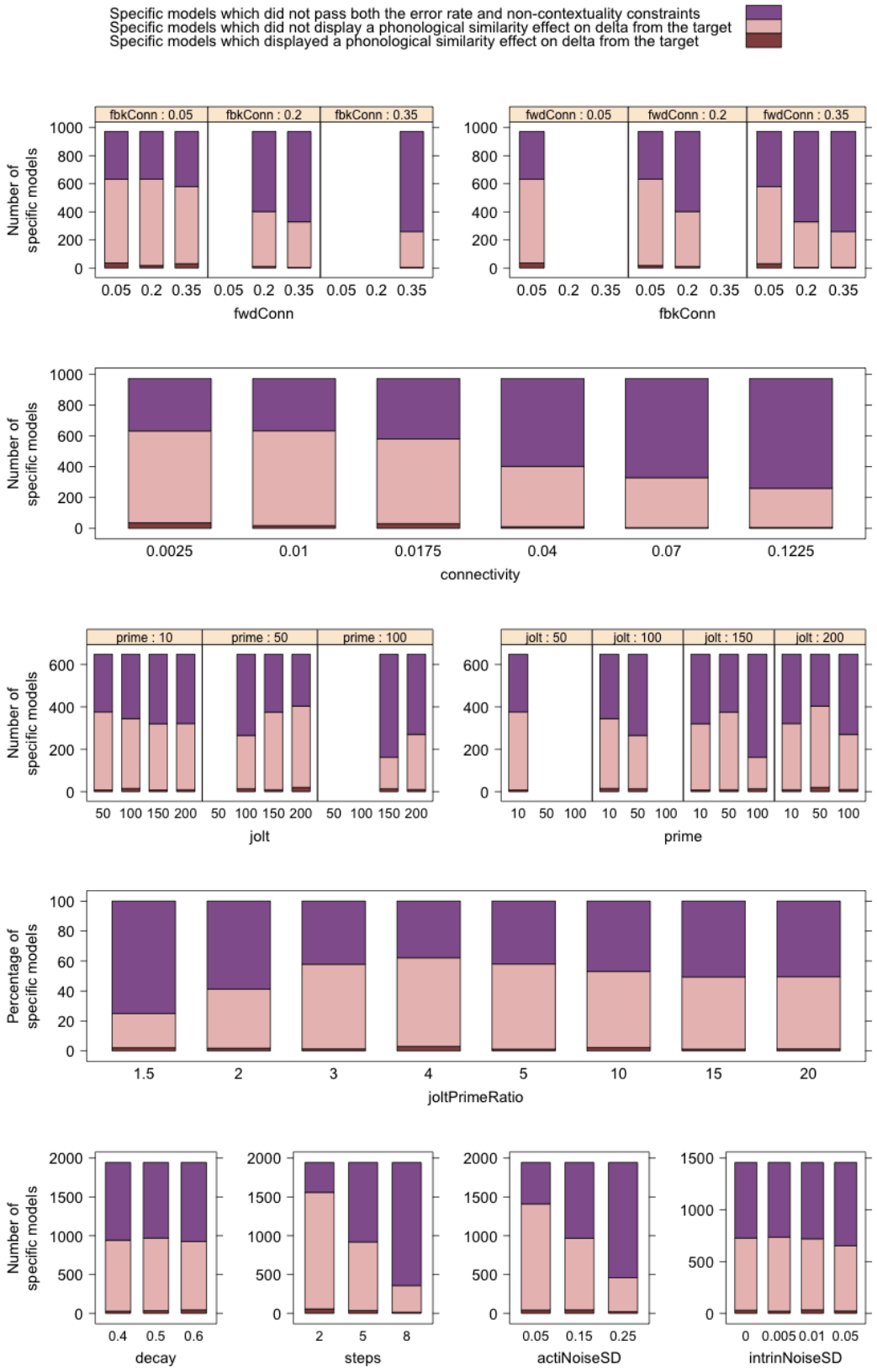


Figure 8.15: The effect of parameter manipulations on whether two-stage models with no feedback from phonemes to words and feedback from features to phonemes display a larger delta measured from the target reference for stimuli where the competing onset was similar, for simulated VOT values, with specific models that do not pass both constraints on error rate and non-contextuality of errors marked separately.

Table 8.14: Results of logistic regression model analyses using parameter values to predict whether two-stage models with no feedback from phonemes to words and feedback from features to phonemes display a larger delta measured from the target reference for stimuli where the competing onset was similar, for simulated VOT values. Directions of effects and absolute Wald's Z values are provided, alongside chi-squared test statistics for likelihood ratio tests assessing the contribution of each parameter to the model. An asterisk indicates that a parameter makes a contribution which is significant at the $p < 0.05$ level.

Parameter	Direction	Z	LRT	P (χ^2)	
connectivity	−	0.5	0	0.652	
joltPrimeRatio	−	8.6	73	< .001	*
decay	+	1.9	4	0.057	
steps	+	11.6	137	< .001	*
actiNoiseSD	+	11.8	142	< .001	*
intrinNoiseSD	+	0.8	1	0.446	

features to phonemes and feedback from phonemes to words can explain the finding. The effect on models of the architecture with feedback from features to phonemes and no feedback from phonemes to words appeared to be weak, with very few models exhibiting this behaviour. Correspondingly, once high error rate models were excluded the architecture was unable to account for the evidence. We found no evidence that specific models of any architecture could display a significant effect of phonological similarity on delta for our EPG and ultrasound simulations, regardless of error rate and non-contextuality constraints.

Again, examination of the average feature activation values showed that there were strong frequency effects on both VOT and alveolar and velar readings. In architectures with feedback from features to phonemes but not from phonemes to words, these were primarily effects of feature frequency. Adding feedback from phonemes to words boosted phoneme frequency effects. We further argued that feature activation increases due to frequency are likely to have an even stronger effect on EPG/ultrasound simulations in comparison to VOT simulations, due to the subtraction of one feature value from another which the VOT calculation involves.

We found that significantly more models than would be predicted by chance generated phonological similarity on delta numerical trends for both EPG/ultrasound and VOT simulations for the architecture with feedback from features to phonemes but not from phonemes to words. However, even considering numerical trends we found no evidence of a phonological similarity effect in the architecture with feedback from features to phonemes where feedback from phonemes to words was also present. More worryingly, for some phoneme-to-feature activation flow options,

we found evidence of a reverse phonological similarity trend when feedback from phonemes to words was present. This suggests that there may be a very intricate confound in our well controlled material set, perhaps concerning frequency differences between phoneme pairs and potentially asymmetric effects of these differences in our balanced material set.

Finally, we demonstrated that again, parameters which increase error rate and interactivity support the phonological similarity effect on delta but only to an extent. Too much interactivity results in a reduction of models exhibiting the effect, as a result of frequency effects beginning to swamp the network.

8.4 Conclusions

In this chapter, we aimed to simulate McMillan et al.'s (2009) findings of a lexical bias effect on EPG measurements, and McMillan's (2008) evidence of a phonological similarity effect on EPG, ultrasound and VOT measurements. VOT values were simulated by subtracting the activation value of the voiced feature from the activation value of the voiceless feature as in the previous chapter. The activation levels of the alveolar and velar features were taken to abstractly represent the extent to which the resulting articulation involves tongue raising at the front and the back of the mouth respectively, and were interpreted as simulations of both EPG and ultrasound measurements. It was predicted that all architectures with feedback from phonemes to words would be able to account for McMillan et al.'s (2009) findings regardless of the nature of activation flow between phonemes and features, but that feedback from feature to phonemes would be required to explain McMillan's (2008) phonological similarity results.

However, we found that phoneme and feature frequency had a very large effect on feature activation levels, such that more frequent representations led to higher activation levels for both target and competing features. This caused large amounts of within condition variance, making it harder to detect both lexical bias and phonological similarity effects. The global activation level increase effect of frequency was worse for EPG and ultrasound simulations than it was for VOT simulations, as simulated VOT values were calculated by subtracting one feature activation level from another. Phonological similarity VOT effects were weak due to the effect of frequency, causing problems when the constraints on error rate and non-contextuality were applied, and these frequency effects perhaps help explain why the lexical bias

on VOT traces effect was weak in the previous chapter too. However, when a binomial analysis was carried out to determine whether sufficient models displayed significant effects of lexical bias or phonological similarity on EPG and ultrasound measurements to reject the hypothesis that these effects were due to chance, we found no evidence for significant effects in any architecture at all.

To allow us to investigate whether lexicality and phonological similarity affected these measurements despite the large within condition variance caused by frequency manipulations, we extended the binomial analysis introduced in chapter 6 to determine whether more models showed numerical trends in the predicted direction than could be accounted for by chance. Using this analysis, we demonstrated that as predicted, all architectures with feedback from phonemes to words exhibited a lexical bias effect on EPG measurements. No cascading from phonemes was required.

Using the original binomial analysis of significant effects in models, we showed that phonological similarity effects on VOT measurements analysed with delta could only be accounted for by an architecture with feedback from features to phonemes, and no feedback from phonemes to words. Surprisingly, no evidence was found that the architecture with feedback from features to phonemes and from phonemes to words could explain this result. A binomial analysis of numerical trends gave the same result for VOT measurements, and equally showed that phonological similarity effects on EPG and ultrasound measurements could only be explained by architectures with feedback from features to phonemes but not from phonemes to words. More worryingly, we uncovered a reverse phonological similarity effect in some architectures with feedback from phonemes to words. Whilst our materials were very strictly controlled, as described in section 6.2.2, the only sensible explanation of these results appears to be that our materials contain a very intricate confound of phonological similarity which is related to phoneme frequency (and is therefore only evident when feedback from phonemes to words is present). Such a confound causing a reverse phonological similarity effect would help explain why the architecture with feedback from features to phonemes and from phonemes to words does not display a significant phonological similarity effect in the predicted direction.

Analysis of the effects of parameter manipulations on lexical bias and phonological similarity effects on delta found that parameter settings which supported error generation and activation flow through feedback loops (high connection strengths, high numbers of steps before selection, high levels of activation-based noise and low jolt to prime ratios) increased the likelihood of lexical bias and phonological

similarity effects being witnessed, but only when these parameter settings were not too extreme. For example, at very high connection strengths, the number of models displaying lexical bias and phonological similarity effects began to diminish again. We argued that this reflected the fact that as activation flow increased, frequency effects would grow more quickly than lexical bias and phonological similarity effects, as frequency effects are driven by many nodes whereas lexical bias and phonological similarity effects rely on the presence or absence of just one node. When activation flow increases too much, frequency effects therefore wash out the lexical bias and phonological similarity effects. High decay rates also helped support lexical bias and phonological similarity effects by suppressing activation flow to avoid this frequency effect.

A number of options exist for addressing problems of frequency in the future. Firstly, we could consider building more complicated statistical models than t-tests to evaluate the effect of lexicality and phonological similarity in each specific model. However, we note that the analyses of human data used by McMillan et al. (2009) and McMillan (2008) did not explicitly include frequency. Secondly, these simulations could be repeated in a lexicon where all phonemes have the same frequency. This would not be true to the actual English lexicon however, whereas the randomly selected lexicon used here provided a reasonable model. Thirdly, options for modifying the model to reduce the effects of frequency could be considered, although such a modification would need to allow the lexical bias and phonological similarity effects to remain. Schade and Berg (1992) proposed that laterally inhibitive connections can help address overpowering effects of frequency. However, in their simulations, the remaining effect of frequency is still stronger than the effect of lexical bias. Fourthly, and perhaps most usefully, we could acquire more instrumental evidence as to the effects of frequency so that in future modelling endeavours, there are clearer benchmarks as to what the real effect of these variables should be.

8.5 Chapter summary

In this chapter, we built on the investigations in the previous chapter by modelling EPG as well as VOT evidence, and by modelling instrumental evidence findings which do not rely on categorisation of productions and instead use the *delta method*. We showed that architectures with feedback from phonemes to words demonstrate a lexical bias effect on EPG measurements analysed using the delta method, regardless of the nature of activation flow between phonemes and features. However,

phonological similarity effects on EPG, ultrasound and VOT measurements analysed using the delta method can only be accounted for when feedback from features to phonemes is present.

Phoneme and feature frequency have extremely strong effects on feature activation levels however. Large within condition variance due to lexical bias and phonological similarity effects means that VOT phonological similarity effects are difficult to detect, and EPG and ultrasound effects of lexical bias and phonological similarity only manifest themselves as statistically significant numbers of models displaying numerical trends across architectures, not as statistically significant effects within specific models. Analyses further suggested that despite extremely tightly controlled materials, a complicated reverse confound of phonological similarity relating to phoneme frequency was present in our target phrases. This caused problems for architectures with feedback from phonemes to words, in which no phonological similarity effect was detected, and requires further investigation.

We propose that further instrumental investigations of frequency effects on human articulations would clarify how the behaviour of models of word production should change given manipulations of frequency. Future models are likely to require adjustments to their architecture so that effects of frequency are less overpowering.

Chapter 9

Discussion

9.1 Introduction

In this chapter, we summarise the findings of this thesis and propose a number of ways in which the work presented could be developed.

9.2 Summary of findings

Firstly, we present a synopsis of our results with respect to the behaviour of the original and extended versions of Dell's (1986) model. We then outline the methodological innovations required to obtain these results.

9.2.1 Theoretical findings

Large scale investigation of the behaviour of Dell's (1986) model

The first set of simulations reported in this thesis, in chapters 4 to 6, presented the results of a large scale investigation of the behaviour of Dell's (1986) model in which all eight free spreading activation parameters were orthogonally varied across a range of values representing those previously used in the literature.

We first examined the basic behaviour of Dell's (1986) model at these different parameter settings, considering the overall error rate of the model, and the proportion of errors with a source in nearby words. Of particular interest was our finding that an increase in the number of activation calculation steps before selection results in an increase in error rate. We argue that Dell's (1986) original claims that an increase in the number of steps before selection can be seen as a decrease in speech rate, resulting in reduced error rates, was dependent on model features which are

extremely rare in later implementations of Dell's (1986) framework, and which lead to questionable model behaviour. The number of steps before selection is therefore perhaps better conceptualised as the length of time for which the model must remember the message it intends to convey.

We also highlighted that high error rates can be caused both by low connection strengths, which cause noise in the network to overpower the transmitted signal (Dell, Schwartz, et al., 1997), as well as by high forward or feedback connection strength when feedback connections are present, which can lead the network to accumulate activation and produce utterances which reflect the underlying structure of the network rather than the intended message. This finding is in line with Goldrick's observation (e.g., Goldrick, 2006) that overly strong feedback can lead to behaviour which does not reflect empirical human word production results.

Keeping in mind that a good model of speech error production in humans cannot generate too many errors, we used empirical evidence to calculate upper limits on error rate and the proportion of errors which do not have a source in nearby words. Throughout the thesis, we considered which parameter settings allowed the model to meet these constraints.

We additionally reanalysed speech error corpus reports to determine bounds on the relative proportions of anticipations, perseverations and exchanges produced by humans. Examination of the proportions of these errors generated by Dell's (1986) original model revealed that it can account for the finding that speakers who make more errors also make a higher proportion of perseverations (Dell, Burger, & Svec, 1997), and predicts that speakers who make a higher proportion of perseverations should also make more errors in which the source of the error is outside the current utterance. However, we found that the model could not generate an adequate proportion of exchange errors without exceeding upper bounds on error rates.

Further investigation of the effects of parameter manipulations on this behaviour strongly suggested that parameter settings outside those explored in this simulation would not be able to close the substantial gap between the model behaviour observed in these studies and empirical results. Increasing connection strength and decreasing decay offered most potential. However with a current maximum connection strength of 0.35 and decay rate of 0.4, limited movement is possible given Shrager et al.'s (1987) demonstration that connection strength must be lower than decay rate to prevent activation rising without bound such that the network would not be able to encode messages. In any case, the current results demonstrate that across a very

wide range of parameter settings, encompassing those used in the literature, the model cannot account for the exchange rate results. This result demonstrates how a large scale parameter search approach can offer greater insight into the general behaviour of the underlying architecture. Results later in the thesis consider the behaviour of the model on the first word only in order to excise the problematic word sequencing mechanism.

Finally, we provided statistical evidence that in a model with output at the phoneme level, as in Dell's (1986) original simulations, the lexical bias effect requires feedback from phonemes to words, and the phonological similarity effect requires feedback from features to phonemes. Parameter settings which increase the error rate of the model, such as a low jolt to prime ratio and high levels of activation-based noise, support these effects by increasing the power of the per specific model statistical analysis. Settings which increase activation flow through the network, such as high connection strength and a high number of steps before selection increase the size of these effects by increasing the influence of feedback loops in the network.

Information flow between phonemes and features

In the second set of simulations reported in this thesis, in chapters 7 and 8, we extended Dell's (1986) model of phonological encoding to consider output at a subphonemic, or in our implementation, featural level. We investigated whether this extended model could account for transcribed speech error evidence and new instrumental acoustic and articulatory results, some of which relied on perceptual categorisation of productions as erroneous or correct and some of which did not. In particular we examined what constraints such evidence placed on models of activation flow between phonemes and subphonemic representations.

We considered four models of information flow from phonemes to subphonemic representations: a model with *no cascading from phonemes* in which only the identity of the selected phoneme is conveyed to the subphonemic level; a model with *cascading from selected phonemes only*, where the activation level of the selected phoneme is transmitted to the subphonemic level; a model with *cascading from all phonemes*, in which activation from all phonemes passes to the subphonemic level; and finally a model with *feedback from subphonemic representations*, where activation also feeds back to the phoneme level.

Table 9.1: The ability of different two-stage models of information flow to account for empirical data, according to simulations.

	Results					
	transcribed LB	transcribed PS	G&B 2006 traces	G&B 2006 trace LB	MMea 2009 delta LB	MM 2008 delta PS
Cascading from all Ws to Ps						
No cascading from Ps to Fs	×	✓	✓	×	×	×
Cascading from selected Ps to Fs	×	✓	✓	×	×	×
Cascading from all Ps to Fs	×	✓	✓	×	×	×
Feedback from Fs to Ps	×	✓	✓	×	×	✓
Feedback from Ps to Ws						
No cascading from Ps to Fs	✓	✓	✓	✓	✓	×
Cascading from selected Ps to Fs	✓	✓	✓	✓	✓	×
Cascading from all Ps to Fs	✓	✓	✓	✓	✓	×
Feedback from Fs to Ps	✓	✓	✓	✓	✓	×

Key:
LB = lexical bias, PS = phonological similarity, G&B 2006 = Goldrick and Blumstein (2006), MM 2008 = McMillan (2008), MMea 2009 = McMillan et al. (2009)

Ws = words, Ps = phonemes, Fs = features

✓ = shown to be able to account for empirical results

× = no evidence found for ability to account for empirical results

Grey boxes indicate that results did not match the standard claim in the literature.

Bordered boxes indicate that results did not match predictions in chapter 2.

Our simulations largely focused on results from the literature which made strong claims about phoneme to subphonemic representation information flow. In particular, Goldrick and Blumstein (2006) presented acoustic evidence of an influence of intended phonemes on erroneous productions of other phonemes, and claimed that an account of this result required cascading from all phonemes. They supported this claim with a post-hoc analysis demonstrating a lexical bias on traces of the intended phoneme. McMillan (2008) provided further evidence of phonological similarity effects on acoustic and articulatory measurements, and used these to claim that activation must feed back from subphonemic representations to phonemes.

Our results are summarised in table 9.1. We found that all of the models of activation flow from phonemes to features, including the most discrete model in which only the identity of the phoneme is conveyed to the feature level, can account for most of the simulated data. It was no surprise to find that we could account for the transcribed lexical bias effect in the simplest model when feedback from phonemes

to words was present. Equally, whilst McMillan et al. (2009) presented their results in the framework of a model with cascading from all phonemes, no very strong claims were made that a more discrete model would not also be able to account for this result.

In contrast, Goldrick and Blumstein (2006) presented their results as strong evidence for cascading from all phonemes. These simulations allowed us to demonstrate the validity of our argument that there were two further ways for less interactive models to account for their main finding of influences of intended phonemes on erroneous productions in acoustic recordings. However, we were surprised to find that the most discrete model could also account for a lexical bias effect on these acoustic traces of intended phonemes, whereas we had predicted that at least cascading from selected phonemes would be required. This finding highlighted an oversight in our own careful pen and paper based reasoning about model behaviour.

Within the framework of Dell's (1986) original model, the transcribed phonological similarity effect has been explained by feedback from features to phonemes. However, the simulations in this thesis demonstrated that no feedback from features to phonemes is required to account for this result when output is at the featural level. Our investigations showed that of all the evidence considered here, the only finding placing a constraint on information flow between phonemes and features was McMillan's (2008) demonstration of a phonological similarity effect on acoustic and articulatory measurements, which was only exhibited by models with feedback from features to phonemes.

We note that it may be possible to make further arguments about models of activation flow required to account for this evidence, based on the distributions of VOT and articulatory measurements observed. In this case however, it would be very important to emphasise that it is distributional characteristics of the data which differentiate between the models. Our results have shown that almost none of the statistical differences considered in this thesis can be used for this purpose.

There were two key problems with the models reported in this thesis. Firstly, due to our decision to apply priming at the word level, models with no cascading from unselected phonemes to features, or no feedback from phonemes to words, struggled to generate contextual errors at the featural level, as priming activation either could not reach this level due to limited activation flow, or decayed away before subphonemic processing began due to lack of feedback reinforcement. Models which relied on contextual error generation at the featural level therefore had to

generate very high error rates overall in order for enough data to be available for analysis, as contextual errors were not more common than featural errors of other kinds. Direct priming of phonemic and subphonemic representations may occur however, for example in tongue twisters, or because of perseveratory influences from a recently produced sound. Future models in which such priming was implemented would presumably not experience such problems.

Secondly, feedback loops in the model which drive the lexical bias and phonological similarity effects also cause more frequent representations to become highly activated. When analysing model output based on feature activation rather than feature selection, we found that these frequency effects overpowered lexical bias and phonological similarity effects by causing large within condition variation. The effect of the activation increase throughout the network which was caused by production of frequent representations was reduced for VOT simulations, as calculation of the simulated VOT value within our results relied on subtraction of one feature activation level from another. Effects in simulations of EPG and ultrasound evidence were only detectable by running statistics over the numerical trends displayed by all specific models of a given activation flow architecture, as per specific model statistical effects were generally not significant. Furthermore, we observed a potential reverse confound of frequency related variables with phonological similarity in our material set which exhibited itself in models with feedback from phonemes to words, such that the architecture with both feedback from features to phonemes and feedback from phonemes to words was unexpectedly unable to account for McMillan's (2008) findings. The desired and actual effect of frequency on the model's behaviour needs further consideration, as outlined in section 9.3.1.

Throughout our investigations, we clarified the influence of parameter setting manipulations on the models' ability to account for empirical evidence. As for the model with output at the phoneme level, we found that many effects were more likely to be exhibited when more errors occurred, for example at low jolt to prime ratio settings and where high levels of activation-based noise were present. More errors equated to more data for the statistical analyses, boosting their power. Also in line with our results from the model with output at the phoneme level, we found that transcribed lexical bias effects were more likely to be found when connection strengths and the number of steps before selection were high, such that activation flow through feedback loops was encouraged. However, for simulations of lexical bias and phonological similarity effects on articulatory measurements, evidence was found that settings which provided very strong support for activation flow through

feedback loops caused frequency effects to overwhelm effects of lexical bias and phonological similarity, such that a happy medium was required.

Interestingly, we showed that different parameters were required to account for traces of intended phonemes on erroneous productions (Goldrick & Blumstein, 2006) in the model with no cascading from phonemes in comparison to other models. The no cascading account of trace generation relied on errors at a featural level. Since the featural level in this model did not receive priming support for contextual errors, trace generation was supported by parameter settings which led to very low signal activation levels (low connection strength, high decay rates and high numbers of steps before selection) such that intrinsic noise governed activation levels and overall error rates were high. Trace generation due to errors at the phoneme level, which was the main source of traces in all other models, required entirely contrasting parameters. This mechanism was supported by high forward connection strength, low decay rates and a low number of steps before selection, such that the activation pattern generated on phonemes at selection could be faithfully transmitted to the featural level. Similarly, accounts of phonological similarity which relied on featural errors generally required low activation levels in the network, as driven by low connection strengths, whereas interactive accounts supported by feature-to-phoneme feedback were more successful at high connection strengths. This demonstrates that investigations of behaviour of different information flow options at one arbitrarily chosen set of parameter settings may well have given rise to misleading results.

General theoretical conclusions

The findings presented above raise two general questions about the human word production architecture. Firstly, is it valid to continue to use principles suggested by Dell (1986) when considering the mechanics of word production, or does the model's inability to generate a sufficiently high proportion of exchange errors falsify the architecture completely? Secondly, how do the current results speak to the more general question of whether there is cascading or feedback in the word production system?

We argue that the current results do not mean that we should abandon Dell's (1986) model completely. Implementations of Dell's (1986) architecture which produce single words only have accounted for vast swathes of normal and aphasic speech error production data (e.g. Dell, Schwartz, et al., 1997; Foygel & Dell, 2000; Goldrick, 2006; Rapp & Goldrick, 2000). The implication of the current results is that adding

a mechanism to the model which primes upcoming words and phonemes and resets produced units to allow the model to generate sequences of words, does not allow the model to account for core evidence about errors in word sequences. It is therefore the implementation of this word sequencing mechanism which should be rejected, rather than the entire model.

On a broader note, it appears unlikely that any one piece of evidence would ever falsify the entire Dell (1986) architecture. The model is composed of a wide array of assumptions, such as the nature of representations used, the level at which the output of the model is measured, the nature of information flow between representations and the way in which sequences of units are produced. It is much more likely that these assumptions would be challenged individually, as in the current thesis.

Our current results also allow us to draw some conclusions about the presence of cascading and feedback at certain points in the word production system. Specifically, they strongly suggest that without the implementation of a monitor, feedback from phonemes to words is required to account for lexical bias effects in transcribed and instrumental data. Furthermore, articulatory results of manipulations of onset phoneme similarity (McMillan, 2008) cannot be accounted for without feedback from features to phonemes.

However, a key goal of this thesis has been to clearly identify which pieces of evidence motivate which specific attributes of a model of the word production system, and to be explicit about which assumptions such conclusions are based on. It is therefore important to remain aware that the conclusion that feedback is required from features to phonemes is dependent on one piece of evidence only, and therefore if this evidence was challenged, this claim could no longer be made. Equally, the need for feedback from phonemes to words has only been demonstrated in a model which, like all other current implementations of word production systems, has no implemented monitoring system. If this assumption was changed, the conclusion may also change. Lastly, in the absence of strong evidence for an assumption of uniform information flow throughout the word production architecture, an extension of the conclusions drawn here to parts of the word production system above the word or below the feature would not be productive. Further experimental and computational examination of the information flow in other parts of the word production system, or acquisition of evidence that information flow principles should be common to the system as a whole, would be more beneficial.

This approach of raising awareness of where the limitations to our conclusions lie facilitates the efficient design of future experimental and modelling work. In turn, this encourages speedier progress in the reverse engineering of the human word production system.

A final general conclusion to be drawn from the investigations presented in this thesis relates to the role of modelling in the evaluation of spreading activation theories. Across a range of results we have demonstrated that human pen-and-paper based reasoning, whether carried out by other researchers or by ourselves, can result in false beliefs that a spreading activation model can account for evidence that in reality it cannot, or that such a model cannot account for evidence that in reality it can. The graphical appearance of these spreading activation models is deceptively simple, whilst their true emergent behaviour can be substantially more complex. To ensure maximum velocity towards a full understanding of human word production or other cognitive systems modelled using this framework, it is paramount that conclusions about the ability of spreading activation theories to account for empirical evidence are firmly grounded in simulations over implemented models.

9.2.2 Methodological advances

To uncover the results reported in the previous section, a number of methodological innovations were required. We demonstrated that the behaviour of models within the framework proposed by Dell (1986) can be relatively easily investigated at many parameter settings by making use of cluster computing technology. We presented a simple method for analysing the large amount of data generated by such an approach to reveal the effect of parameter settings on the model's behaviour. We also showed how we can determine whether a particular model of activation flow can account for a given statistical difference demonstrated in human behaviour, when statistical tests of the model are carried out at many different parameter settings, such that there is a very high chance of some false positive results occurring. This methodology was extended with some success to investigate whether architectures can account for multiple effects without requiring a change of parameter settings, but future research should investigate how the multiple effect analysis can achieve greater power.

9.3 Future work

In this section, we consider possible theoretical and methodological directions for future work.

9.3.1 Theoretical directions

The work presented here highlights a number of opportunities for theoretical development. In the first part of the thesis where we considered the behaviour of Dell's (1986) original model, we demonstrated that there is a need for a new model of exchange error generation. We proposed for example that the effect of repeated post-selection inhibition of phonemes could be investigated. An ideal new model of word sequencing may simultaneously address the lack of within-word sequencing in Dell's (1986) model and its inability to explain the prevalence of onset errors, as highlighted in chapter 3.

In the second part of the thesis, in which we considered the behaviour of a model with output at the featural level, some architectures with limited activation flow experienced problems due to priming activation being applied at the word level, such that there was no priming support for contextual error generation at the featural level. Parameter settings which led these models to generate sufficient contextual errors for analysis also caused high overall error rates due to the large number of non-contextual errors generated. We have argued that applying priming at a phonemic and subphonemic level would allow models to display the desired effects at much lower overall error rates. It would of course be useful to run simulations in which this suggestion was implemented, in order to verify that this this model modification did not cause undesired repercussions for other aspects of model behaviour.

Simulations presented in the second part of the thesis were subject to strong effects of frequency on feature activation levels, which outweighed desired effects of lexical bias and phonological similarity. We have noted that further instrumental investigations of the effect of these frequency variables would provide better benchmarks for future modelling endeavours. However, current results would suggest that a reduction of the frequency effect relative to lexical bias and phonological similarity effects may be required. Schade and Berg (1992) have suggested that lateral inhibition can help reduce effects of frequency, which caused problems in simulations of lexical bias and phonological similarity effects on articulatory and acoustic measures, although it is not clear whether lateral inhibition would also reduce the lexical bias and phonological similarity effects themselves. Interestingly however, Harley

(1993) presents results suggesting that in a model with lateral inhibition, competitor activation decreases as timesteps pass, whilst target activation remains high. Presumably error rate in such a model would therefore decrease as timesteps passed, as Dell (1986) originally reported, instead of error rate increasing as timesteps pass as the simulations reported in this thesis show. Further investigation of lateral inhibition models would therefore potentially be informative. Otherwise, it would be useful to find a mechanism for reducing frequency effects, which are driven by connections to many nodes, in relation to lexical bias and phonological similarity effects, which are driven by connections to a single node.

It would also be interesting to try to extend the simulation work presented here to consider other effects of high level variables which have been demonstrated on articulatory and acoustic measures. For example, a number of experimental results have demonstrated effects of word frequency and neighbourhood effects on aspects of articulation, ranging from the voicing of stops (Baese-Berk & Goldrick, 2009) to vowel space and duration (Munson, 2007; Munson & Solomon, 2004; Wright, 2004). Pouplier and colleagues (Goldstein et al., 2007; Pouplier, 2007) have also demonstrated the occurrence of gestural intrusions (e.g., where the tongue dorsum raises during production of an alveolar consonant to over two standard deviations more than the control mean height for alveolar consonants), and have reported that these are more frequent than gestural reductions (e.g., where the tongue tip lowers during production of an alveolar consonant to over two standard deviations less than the control mean height for alveolar consonants). This evidence has been presented in favour of the gestural model of articulation (Browman & Goldstein, 1992). Simulations of these results would clarify whether we can continue to extend Dell's (1986) model to account for further articulatory results, and may help determine further constraints on models of activation flow between phonemes and subphonemic representations. Finally, where evidence of manipulations of certain variables does not exist in the experimental literature, it would be interesting to manipulate these variables within the model itself to investigate what predictions arise.

9.3.2 Methodological directions

There are a number of ways in which the methodology presented in this thesis could be further developed.

Firstly, we presented a binomial analysis to allow us to determine whether models are able to account for a statistical pattern reported in empirical investigations, when they have been tested at multiple parameter settings. An extension of this method was also outlined and applied, to allow us to determine whether models could account for multiple statistical patterns without requiring a change in parameter settings. The current version of this extended analysis experiences a loss of power as more statistical patterns must be accounted for. It would be useful if future work could address this limitation.

Secondly, nearly all of the simulations presented in this thesis focus on accounting for previous results. It would also be possible to use the large scale modelling approach to generate model behaviour predictions, which could be further specified in terms of parameter settings required for a given pattern. Future work could demonstrate and elaborate on such a prediction methodology.

Thirdly, all simulations in this thesis have assumed that parameters at different parts of the model have the same value. For example, forward connection strength from words to phonemes was always equal to forward connection strength from phonemes to features. There is no evidence to require that this is the case however. Future investigations could therefore consider model behaviour with separate parameter settings for separate stages (see, e.g., Foygel & Dell, 2000, for a model of aphasic patients in which connection strength differs for different stages).

Lastly, when modelling speech error results, there are conflicting constraints in that the model must generate speech errors in order for there to be sufficient data for analysis, but models which generate high error rates are not appropriate models of human production. It may in future be worth considering running an even higher number of trials per specific model, to allow low error rate simulations a greater opportunity to exhibit significant speech error effects.

9.4 Conclusions

Our simulations showed that Dell's (1986) account of word sequencing, as implemented by the priming, selection and check off mechanisms, can account for negative correlations between the proportion of movement errors which are anticipatory, and error rate (Dell, Burger, & Svec, 1997). However, we demonstrated that corpus evidence of speech errors is strongly out of line with the exchange error generation behaviour of the model, such that the model produces far too few exchanges, or

far too many errors overall. We are therefore forced to conclude that the word sequencing account in its current form is deficient.

The spreading activation nature of the model can to a large extent be considered independently of the word sequencing account, however. In this thesis, we showed that this model feature not only allows the model to account for a wide range of transcribed speech error findings, such as the lexical bias and phonological similarity effect, but also permits it to explain new articulatory and acoustic evidence.

We examined how interactive the activation flow between phonemes and features must be to account for new evidence, or in other words, to what extent activation must cascade from phonemes to features, and whether it feeds back. Whilst simulations verified that these new articulatory and acoustic findings provided further evidence for feedback between phonemes and words (Goldrick & Blumstein, 2006; McMillan et al., 2009) we found that a very discrete account of activation flow between phonemes and features can account for a number of new instrumental findings. In particular, our simulations showed that to account for results presented as strong evidence for cascading from all phonemes (Goldrick & Blumstein, 2006), no cascading from phonemes is in fact required. However, whilst we demonstrated that the transcribed phonological similarity effect is not evidence for feedback from features to phonemes in a model with output at the feature level, our results appear to confirm that such feedback is required to explain McMillan's (2008) findings of a phonological similarity effect on instrumentally acquired acoustic and articulatory data.

Feedback not only drives the lexical bias and phonological similarity effects however, but also causes frequency effects, providing particular activation to representations with a high number of connections to other representations. In our instrumental results, frequency effects were stronger than lexical bias and phonological similarity effects, such that the influence of lexical bias and phonological similarity was harder to detect. Whilst further investigation of the effect of frequency on instrumental measurements is required to establish better benchmarks for the model, it is possible that a modification of the model will be necessary, such that the lexical bias and phonological similarity effects as driven by feedback from a single node continue to be evident, but frequency effects as driven by feedback from multiple nodes are reduced.

In the future, it would be interesting to extend this work to other instrumental results (e.g., Baese-Berk & Goldrick, 2009; Goldstein et al., 2007) to see if they can

also be accounted for, and to clarify what constraints they place on information flow between phonemes and features. Manipulation of further variables in the model would allow predictions to be generated for empirical testing.

This thesis has not only provided theoretical insight into Dell's (1986) model of word production and an extension of this model which accounts for new instrumental evidence, but has provided further illustration of the value of modelling, and in particular, the advantages of a large scale parameter varying approach. For example, this approach allowed us to demonstrate that the model cannot generate a sufficiently high proportion of exchange errors. By examining exchange error generation behaviour at many parameter settings, we were able to increase our understanding of the behaviour of the underlying architecture. Results showing that different parameter settings were required for trace generation (Goldrick & Blumstein, 2006) in different architectures confirmed that an approach investigating the behaviour of different models of information flow at a single set of arbitrarily chosen parameters settings is inappropriate. We further emphasised the general need for explicit modelling of theories by finding holes in our own pen-and-paper reasoning and demonstrating that a lexical bias on traces (Goldrick & Blumstein, 2006) can be accounted for without any cascading from phonemes to features. Where predictions were shown to be correct, further insight into the necessary characteristics of a model to account for this evidence was provided by an examination of the parameter settings necessary for such model behaviour. We argue that the advances in technology which the past 20 years have brought have gradually removed all excuses to ignore the parameter settings in Dell's (1986) model.

References

Agresti, A. (2002). *Categorical data analysis*. Hoboken, New Jersey: John Wiley & Sons Inc.

Baars, B. J., & Motley, M. T. (1976). Spoonerisms as sequencer conflicts: Evidence from artificially elicited errors. *American Journal of Psychology*, *89*, 467–484.

Baars, B. J., Motley, M. T., & MacKay, D. G. (1975). Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behaviour*, *14*, 382–391.

Baese-Berk, M., & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and Cognitive Processes*, *24*, 527–554.

Boucher, V. J. (1994). Alphabet-related biases in psycholinguistic inquiries: Considerations for direct theories of speech production and perception. *Journal of Phonetics*, *22*, 1–18.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*, 201–251.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica*, *49*, 155–180.

Buckingham, H., & Yule, G. (1987). Phonemic false evaluation: Theoretical and clinical aspects. *Clinical Linguistics & Phonetics*, *1*, 113–125.

Cole, R. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception & Psychophysics*, *13*, 153–156.

Crompton, A. (1981). Syllables and segments in speech production. *Linguistics*, *19*, 663–716.

Cutler, A. (1981). The reliability of speech error data. *Linguistics*, *19*, 561–582.

del Viso, S., Igoa, J. M., & Garcia-Albea, J. E. (1991). On the autonomy of phonological encoding: Evidence from slips of the tongue in Spanish. *Journal of Psycholinguistic Research*, *20*, 161–185.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283–321.

Dell, G. S. (1988). The retrieval of phonological forms in production: Test of predictions from a connectionist model. *Journal of Memory and Language*, *27*, 124-142.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, *5*, 313–349.

Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: a functional analysis and a model. *Psychological Review*, *104*, 123–147.

Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? In A. S. Meyer & N. O. Schiller (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 9–37). Berlin: Mouton de Gruyter.

Dell, G. S., Lawler, E. N., Harris, H. D., & Gordon, J. K. (2004). Models of errors of omission in aphasic naming. *Cognitive Neuropsychology*, *21*, 125–145.

Dell, G. S., & O'Seaghdha, P. G. (1991). Mediated and convergent lexical priming in language production: a comment on Levelt et al. (1991). *Psychological Review*, *98*, 604–14.

Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, *42*, 287–314.

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior*, *20*, 611–629.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*, 801–838.

Fay, D., & Cutler, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry*, *8*, 505–520.

Ferreira, V., & Griffin, Z. (2003). Phonological influences on lexical (mis) selection. *Psychological Science*, 86–90.

Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, *43*, 182–216.

Frisch, S. A., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics*, *30*, 139–162.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, *47*, 27–52.

Fromkin, V. A. (Ed.). (1973). *Speech errors as linguistic evidence*. The Hague: Mouton.

Gagnon, D. A., Schwartz, M. F., Martin, N., Dell, G. S., & Saffran, E. M. (1997). The origins of formal paraphasias in aphasics' picture naming. *Brain and Language*, *59*, 450–472.

Garnham, A., Shillcock, R., Brown, G. D. A., Mill, A. I. D., & Cutler, A. (1981). Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics*, *19*, 805-817.

Garrett, M. F. (1975). The analysis of sentence productions. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 9, pp. 133–177). New York: Academic Press.

Goldrick, M. (2006). Limited interaction in speech production: Chronometric, speech error, and neuropsychological evidence. *Language and Cognitive Processes*, *21*, 817–855.

Goldrick, M., & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, *21*, 649–683.

Goldrick, M., & Rapp, B. (2002). A restricted interaction account (RIA) of spoken word production: The best of both worlds. *Aphasiology*, *16*, 20–55.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition*, *103*, 386–412.

Harley, T. A. (1984). A critique of top-down independent levels models of speech production: Evidence from non-plan-internal speech errors. *Cognitive Science*, *8*, 191–219.

Harley, T. A. (1993). Phonological activation of semantic competitors during lexical access in speech production. *Language and Cognitive Processes*, *8*, 291–309.

Hartsuiker, R. J. (2002). The addition bias in Dutch and Spanish phonological speech errors: The role of structural context. *Language and Cognitive Processes*, *17*, 61–96.

Hartsuiker, R. J., Corley, M., & Martensen, H. (2005). The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related beply to Baars et al. (1975). *Journal of Memory and Language*, *52*, 58–70.

Humphreys, K. R. (2002). *Lexical bias in speech errors.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior (the Hixon symposium).* New York, NY: Wiley.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*, 41–104.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral & Brain Sciences*, *22*, 1–75.

Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., & Pechmann, T. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review*, *98*, 122-142.

Levitt, A. G., & Healy, A. F. (1985). The roles of phoneme frequency, similarity, and availability in the experimental elicitation of speech errors. *Journal of Memory and Language*, *24*, 717–733.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–460.

MacKay, D. G. (1970). Spoonerisms: the structure of errors in the serial order of speech. *Neuropsychologia*, *8*, 323–350.

MacKay, D. G. (1971). Stress pre-entry in motor systems. *American Journal of Psychology*, *84*, 35–51.

MacKay, D. G. (1987). *The organization of perception and action: A theory for language and other cognitive skills.* Springer-Verlag.

Maclay, H., & Osgood, C. (1959). Hesitation phenomena in English. *Word*, *15*, 19–44.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*, 29–63.

Martin, N., Dell, G. S., Saffran, E. M., & Schwartz, M. F. (1994). Origins of paraphasias in deep dysphasia: testing the consequences of a decay impairment to an interactive spreading activation model of lexical retrieval. *Brain Lang*, *47*, 609–660.

Martin, N., Gagnon, D., Schwartz, M., Dell, G., & Saffran, E. (1996). Phonological facilitation of semantic errors in normal and aphasic speakers. *Language and Cognitive Processes*, *11*, 257–282.

Martin, N., Weisberg, R., & Saffran, E. (1989). Variables influencing the occurrence of naming errors: Implications for models of lexical retrieval. *Journal of Memory and Language*, *28*, 462–485.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*, 375–407.

McMillan, C. T. (2008). *Articulatory evidence for interactivity in speech production.* Unpublished doctoral dissertation, University of Edinburgh.

McMillan, C. T., Corley, M., & Lickley, R. (2009). Articulatory evidence for feedback and competition in speech production. *Language and Cognitive Processes*, *24*, 44-66.

Meringer, R. (1908). *Aus dem Leben der Sprache: Versprechen; Kindersprache, Nachahmungstrieb*. Berlin: Behr's Verlag.

Meringer, R., & Mayer, K. (1895). *Versprechen und verlesen [Slips of the tongue and errors in reading]*. Stuttgart: Goeschensche.

Mowrey, R. A., & MacKay, I. R. (1990). Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America*, *88*, 1299–1312.

Munson, B. (2007). Lexical access, lexical representation, and vowel articulation. In J. Cole & J. Hualde (Eds.), *Laboratory phonology 9* (p. 201-228). New York: Mouton de Gruyter.

Munson, B., & Solomon, N. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language and Hearing Research*, *47*, 1048.

Nooteboom, S. G. (1969). The tongue slips into patterns. In A. G. Sciarone, A. J. van Essen, & A. A. van Raad (Eds.), *Nomen: Leyden studies in linguistics and phonetics* (pp. 114–132). The Hague, The Netherlands: Mouton.

Nooteboom, S. G. (1980). Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 87–95). London, UK: Academic Press.

Nooteboom, S. G. (2005a). Lexical bias revisited: Detecting, rejecting and repairing speech errors in inner speech. *Speech Communication*, *47*, 43–58.

Nooteboom, S. G. (2005b). Listening to oneself: Monitoring in speech production. In R. Hartsuiker, R. Bastiaanse, A. Postma, & F. Wijnen (Eds.), *Phonological encoding and monitoring in normal and pathological speech* (pp. 167–186). Hove, UK: Psychology Press.

Oppenheim, G. M., & Dell, G. S. (2008). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition*, *106*, 528–537.

Pérez, E., Santiago, J., Palma, A., & O'Seaghdha, P. (2007). Perceptual bias in speech error data collection: Insights from Spanish speech errors. *Journal of Psycholinguistic Research*, *36*, 207–235.

Peterson, R. R., & Savoy, P. (1998). Lexical selection and phonological encoding during language production: Evidence for cascaded processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 539–557.

Pouplier, M. (2007). Tongue kinematics during utterances elicited with the SLIP technique. *Language and Speech*, *50*, 311–341.

Pouplier, M. (2008). The role of a coda consonant as error trigger in repetition tasks. *Journal of Phonetics*, *36*, 114–140.

R Development Core Team. (2008). *R: A language and environment for statistical computing.* Vienna, Austria. (ISBN 3-900051-07-0)

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review*, *107*, 460–499.

Robinson, A. (n.d.). *British English Example Pronunciation (BEEP) dictionary.* Retrieved from World Wide Web: ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*, 60–94.

Ruml, W., & Caramazza, A. (2000). An evaluation of a computational model of lexical access: comment on Dell et al. (1997). *Psychological Review*, *107*, 609–634.

Ruml, W., Caramazza, A., Capasso, R., & Miceli, G. (2005). Interactivity and continuity in normal and aphasic language production. *Cognitive Neuropsychology*, *22*, 131–168.

Ruml, W., Caramazza, A., Shelton, J. R., & Chialant, D. (2000). Testing assumptions in computational theories of aphasia. *Journal of Memory and Language*, *43*, 217–248.

Schade, U., & Berg, T. (1992). The role of inhibition in a spreading-activation model of language production: II. The simulational perspective. *Journal of Psycholinguistic Research*, *21*, 435–462.

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and Language*, *54*, 228–264.

Schwartz, M. F., Saffran, E. M., Bloch, D. E., & Dell, G. S. (1994). Disordered speech production in aphasic and normal speakers. *Brain and Language*, *47*, 52–88.

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 295–342). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Shattuck-Hufnagel, S., & Klatt, D. (1979). The role of word onset consonants in speech production planning: New evidence from speech error patterns. In E. Keller & M. Gupnik (Eds.), *Motor and sensory processes of language* (pp. 17–51). Hillsdale, New Jersey: Erlbaum.

Shrager, J., Hogg, T., & Huberman, B. A. (1987). Observations of phase transitions in spreading activation networks. *Science*, *236*, 1092–1094.

Stemberger, J. P. (1985). An interactive activation model of language production. In A. W. Ellis (Ed.), *Progress in the psychology of language* (pp. 143–186). London: Erlbaum.

Stemberger, J. P. (1989). Speech errors in early child language production. *Journal of Memory and Language*, *28*, 164–188.

University of Edinburgh. (2007, August). *Edinburgh Compute and Data Facility web site.* http://www.ecdf.ed.ac.uk/.

Vitevitch, M. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 735–747.

Vousden, J. I., Brown, G. D. A., & Harley, T. A. (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology*, *41*, 101–175.

Waltz, D. L., & Pollack, J. (1985). Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science*, *9*, 51–74.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *176*, 392–393.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior And Development*, *7*, 49–63.

Wilshire, C. E. (1999). The "tongue twister" paradigm as a technique for studying phonological encoding. *Language and Speech*, *42*, 57–82.

Wright, R. A. (2004). Factors of lexical competition in vowel articulation. In R. O. J. J. Local & R. Temple (Eds.), *Laboratory phonology* (Vol. 6, pp. 26–50). Cambridge, UK: Cambridge University Press.